Educational Measurement

EDUCATIONAL MEASUREMENT

Educational Measurement

E. F. LINDQUIST

EDITOR

WITH CHAPTERS BY

GORDON V. ANDERSON
HENRY CHAUNCEY
HERBERT S. CONRAD
WALTER W. COOK
EDWARD E. CURETON
JOHN G. DARLEY
FREDERICK B. DAVIS
ROBERT L. EBEL
JOHN C. FLANAGAN
NORMAN FREDERIKSEN
E. F. LINDQUIST
IRVING LORGE
CHARLES I. MOSIER
DAVID G. RYANS
GERALDINE SPAULDING
JOHN M. STALNAKER
ROBERT L. THORNDIKE
ARTHUR E. TRAXLER
RALPH W. TYLER
K. W. VAUGHN

AMERICAN COUNCIL ON EDUCATION
WASHINGTON, D. C.

The preparation of this manuscript and its publication were made possible by the grant of funds by The Grant Foundation, Inc., New York

Published February 1951
Second printing, January 1955

PRINTED IN THE UNITED STATES OF AMERICA
BY GEORGE BANTA PUBLISHING COMPANY, MENASHA, WISCONSIN

FOREWORD

It is well known that the measurement of individual ability, achievements, and characteristics offers the most solid basis on which students may be assisted in their choice of studies and occupations. Although individual measurement was once regarded with natural suspicion, research in this field has made such rapid progress as now to command the respect and confidence of personnel officers both in schools and colleges, on the one hand, and in industry, on the other. The movement may, indeed, now be regarded as having established itself as the chief source of information on which educational and personnel officers may rely to aid them in their processes of selection and guidance of individuals.

In the belief that the field of individual measurement offered great possibilities for improving educational processes, the American Council on Education has long given continuous and close attention to various aspects of the movement. For many years it sponsored the American Council Psychological Examination. The Council founded the Cooperative Test Service and the National Teacher Examinations, as well as numerous other enterprises in the testing field.

The Council has now turned over the operation of these activities to the Educational Testing Service, which in 1948 it helped to establish jointly with the College Entrance Examination Board and the Carnegie Foundation for the Advancement of Teaching.

The Council retains full liberty, however, to carry on studies concerning the areas in which there is special need for new instruments of evaluation. Indeed, continuous appraisal of the effectiveness and usefulness of testing instruments should remain primarily the responsibility of educational organizations.

Certainly, developments in the testing field are both so extensive and complex as to justify and make necessary comprehensive periodical reviews in order that we may know where we are. For this reason it is particularly appropriate that the Council's Committee on Measurement and Guidance should have planned the general character and manner of executing the publication of the present book on educational measurement.

It is a pleasure as well as a privilege to make acknowledgment to The Grant Foundation for the subsidy which made this project possible, and to Mr. W. T. Grant personally for his understanding of and interest in the study of the individual. Mr. Grant's interest in extending the scope of personality analysis has been a gratifying encouragement to the American

Council on Education and to everyone who has collaborated in the production of this volume.

It should be pointed out that the subsidy was used exclusively to bear the clerical and incidental costs in the preparation of the manuscript (including the expenses of the planning and editorial conferences) and the costs of publication. The work of the writers, collaborators, and the editor was contributed without compensation as a professional service. No royalties to individuals are to be paid out of the proceeds from the sale of the volume. All income from sales is to remain in a permanent Measurement Book Fund of the American Council on Education, to be spent on future revisions of the volume and on the preparation and publication of supplementary brochures on special topics.

The Council is deeply indebted to each and every person who has taken the time to contribute to the planning and completion of this volume, but especially to Dr. E. F. Lindquist for the sound scholarship, patient devotion, and broad point of view which he has brought to his duties as editor. I believe that the volume will immediately establish itself as a landmark in the development of the testing movement.

GEORGE F. ZOOK, *President*
American Council on Education

July 1950

EDITOR'S PREFACE

THERE HAS LONG BEEN AN URGENT NEED, PARTICULARLY IN THE advanced training of measurement workers, for a comprehensive handbook and textbook on the theory and technique of educational measurement. At the time that this volume was planned, in 1945, no book had yet been published that would even begin to fill this need. Anyone then attempting to offer a course in the theory and technique of educational measurement at the advanced graduate level found it almost impossible to provide the student with adequate reference and instructional materials. Much that is of real value had been written and published in this field, but most of it was virtually inaccessible to students in general and particularly to field workers. Most of it had appeared in articles which were very widely scattered in a large number of different periodicals over a period of many years, while the rest had appeared in special bulletins of various testing agencies, in committee reports and in reports of conference proceedings, in test manuals, and in other sources that had never obtained general distribution and were often unavailable even in the best of university libraries.

Not only was most of what had been written not readily accessible to students and field workers, but much of what needed to be made available to them had never been written up and published at all. Most of the published articles were concerned with specific contributions to measurement theory, particularly the mathematical theory, or with some newly developed technique, or with a specific research study. There were few, if any, articles concerned with what might be described as the *art* of test construction or of item writing, essays on the underlying philosophy of educational measurement, articles in which the writer attempted to summarize, evaluate, or elucidate the contributions of others, or articles in which he tried to pass on to others the detailed knowledge, the "tricks of the trade," and the improved sense of values that he had acquired through practical experience. Judging by past production, this was a type of writing which would be done only in the preparation of a textbook, or on definite assignment as a professional obligation.

Primarily, perhaps, as a result of this lack of reference and instructional materials, there were in 1945 very few educational institutions in which any systematic courses in educational measurement were being offered at an advanced graduate level; and in the few university courses of this character which were being offered, the instruction undoubtedly fell far short

of its possible effectiveness, for the reasons given. It appeared, therefore, that until a comprehensive and teachable book in this field was made available, not only would the improvement of educational measurement practice in general be retarded because of inadequate training facilities, but there was danger also that much of what already had been learned would be lost through failure to record it, and would have to be rediscovered and relearned through experience by individual workers.

One educational agency that long had been aware of this situation and anxious to do something about it was the American Council on Education. The Council's standing Committee on Measurement and Guidance had earlier (1936) sponsored a more elementary book of this character, *The Construction and Use of Achievement Examinations,* edited by Hawkes, Lindquist, and Mann, and had since periodically reviewed the need for a similar book at a much more advanced and more technical level. The suggestion was made that the best way for the committee to get this job done was to induce some one individual singlehandedly to prepare the kind of volume needed, and to provide him with the best possible facilities for doing the job. It was contended that only in this manner could a really well-integrated and effective treatment be prepared. The most telling reply to this suggestion was, of course, that in over twenty-five years of the objective testing movement, no one had yet tackled this job or announced his intention of doing so. Because of the ever-increasing degree of specialization within the whole field of measurement, the chances seemed more remote than ever that the volume needed would be produced in this way. Aside from any other possible disqualifications, no one individual had had sufficiently varied practical experience to write authoritatively in all of the various areas of specialization, even if he had the courage and time to tackle so large an assignment. It seemed clear to the committee, therefore, that the kind of book needed could be prepared only through the collaboration of a large number of specialists, each writing in the area of his own special competence.

Accordingly, as soon as the end of World War II permitted, the Measurement and Guidance Committee[1] authorized the calling of a special conference to plan and initiate a project of the general character just suggested. This conference was held in Williamsburg, Virginia, April 17–19, 1945. The members of the planning conference were W. W. Cook, John Flanagan, Irving Lorge, T. R. McConnell, Phillip Rulon, Donald J. Shank, K. W. Vaughn, Ben D. Wood, and E. F. Lindquist, chairman.

[1] The membership of the committee at this time consisted of: Sarah G. Blanding, Galen Jones, E. F. Lindquist, Malcolm Price, E. G. Williamson, Donald J. Shank, Francis L. Bacon, C. L. Cushman, Eugene R. Smith, and George F. Zook.

At this conference the purpose and scope of the projected volume were determined, a tentative table of contents was prepared, writers and collaborators for the various chapters were nominated, and a general editorial policy was set.

In selecting the contributors to the volume, the planning group felt that no one individual alone should be entrusted with the preparation of any single chapter, but that different points of view and types of experience should be represented in every case. It was hoped that each chapter might represent the joint product of the half-dozen or so persons generally recognized as the most competent and experienced thinkers in the area involved, that one of these would take primary responsibility for writing the chapter, and that the others would serve as collaborators, whose duty it would be to review, criticize, and revise what he had written so as to make it generally acceptable. A total of approximately seventy measurement workers were nominated for participation—twenty as chapter writers and the rest as collaborators.

The plans developed at the Williamsburg conference were quickly approved by the Measurement and Guidance Committee and by the American Council on Education, the editor of the projected volume was authorized to proceed at once with execution of these plans, and a generous grant of $20,000 was later obtained by the American Council on Education from The Grant Foundation to pay the expenses of the project.

The project was initially very well received—practically everyone who was invited to participate agreed to do so—but the unusual circumstances existing in the universities following the war prevented many of the contributors from completing their assignments until long past the scheduled time. The original plans had called for completion of the entire project in less than two years, but almost four years passed before first drafts of all of the chapters were in the hands of the editor.

A large share of the editorial work involved was done in a special editorial "work conference" held at Gloucester, Massachusetts, from August 14–27, 1949. The members of this conference were Dorothy Adkins, W. W. Cook, E. E. Cureton, Frederick B. Davis, John Flanagan, Irving Lorge, Charles Mosier, and E. F. Lindquist. Very special acknowledgments are due the members of this conference. For two full weeks, working in editorial teams of two or three per chapter, they collectively read carefully, criticized, and revised more than 1,100 pages of manuscript, and prepared specific suggestions to the writers for further revisions. In several instances, their efforts resulted in major improvements in the original drafts, and many minor inconsistencies, errors, and infelicities were eliminated. Following the Gloucester meeting, furthermore, several members of the

conference contributed a considerable amount of their time in further criticism and revision of the manuscript.

Except in the case of two chapters, the members of the editorial conference and the chapter writers were able to reach essentially full agreement on all important issues considered. In the published versions of these two chapters, opinions are expressed to which the editor and certain members of the editorial conference had objected rather strenuously, but to no avail. This statement is made here, not in criticism of the writers involved, but to explain certain inconsistencies among chapters, and to warn the reader of this volume that he should not assume that authorities are fully agreed on all ideas expressed therein.

The original plans provided for a book of four major parts. The first three of these constitute the present volume. The fourth part was to have contained chapters on Local Testing Programs and Regional Testing Programs, by Max D. Engelhart and E. F. Lindquist, respectively. These chapters were completed early, but had to be omitted from the published volume to keep its size within practicable limits. It is hoped that these chapters may be published later as supplementary brochures.

Major acknowledgment is due to Dr. Ben D. Wood. As director of the Cooperative Test Service from its beginning until 1940, Dr. Wood was an ex officio member of the Measurement and Guidance Committee of the American Council on Education, and it was primarily at his suggestion and insistence that the committee sponsored both the present volume and the earlier volume by Hawkes, Lindquist, and Mann. It was largely through the efforts and intermediation of Dr. Wood, furthermore, that funds were obtained from The Grant Foundation to make this project possible. Certainly it was primarily because of Dr. Wood's personal influence and encouragement that the editor had the temerity to tackle this project and the determination to see it through to completion. This volume might well be regarded as a monument to Dr. Wood's outstanding leadership in educational measurement in this country during the past three decades.

It is obviously impracticable to attempt here to acknowledge individually and adequately all of the many other persons who helped make this volume possible. The names of authors and collaborators are listed at the beginning of the chapters. Everyone concerned—collaborators, conference and committee members, and others—cooperated fully and wholeheartedly in a fine professional spirit, and the editor wishes to express here his sincere personal appreciation of their very generous cooperation.

As stated in Dr. Zook's Foreword, the American Council on Education has established a permanent Measurement Book Project Fund, into which will be paid the proceeds from the sale of this volume. This fund will be used

for future revisions of this volume, and for a series of brochures dealing with special measurement problems which are not of sufficiently general interest to be considered in the present volume. Readers and users of this volume are earnestly requested to cooperate in this continuing project by submitting criticisms and suggestions for revisions of this volume as well as suggested topics for special brochures. It is hoped that with the cooperation of the members of the profession, this project may render increasingly valuable service in the training of measurement workers and in the improvement of educational measurement practices in general.

E. F. LINDQUIST

Iowa City, Iowa
September 1950

CONTENTS

Part Three

MEASUREMENT THEORY

ILLUSTRATIONS

LIST OF TABLES

Part One

THE FUNCTIONS OF MEASUREMENT IN EDUCATION

1. The Functions of Measurement in the Facilitation of Learning

By WALTER W. COOK
University of Minnesota

COLLABORATORS: William A. McCall, *Teachers College, Columbia University;* Herschel T. Manuel, *University of Texas;* Jacob S. Orleans, *The City College of New York;* Ralph W. Tyler, *University of Chicago*

INSTRUMENTS OF EDUCATIONAL MEASUREMENT ARE SIMPLY THE MEANS by which quantitative aspects of human behavior are observed with greater accuracy. To the extent that such instruments conform to the principles of quantitative logic, it becomes possible to know with greater exactness the relationships among the various aspects of educational procedure, the aptitudes of learners, and changes in human behavior. The purpose of this is to make possible more accurate prediction and control in the educational process. The value of measurement depends upon the extent to which the relationships established are crucial from the social point of view. The central questions are: What changes in behavior are desirable? How can these changes be measured? What aptitudes are essential to the development of a given form and level of behavior? What are the crucial elements in the educative process? The value of educational measurement depends upon the validity of the answers to these questions.

Although this book is primarily concerned with the more technical aspects of educational measurement and test construction, it is desirable to give early attention to functions, since instruments are designed and evaluated in terms of their functions. It should be recognized, however, that to state the functions of measurement in other than general terms is somewhat presumptuous. The specific uses of an educational measuring device are limited largely by the ingenuity and insight of the designer and user. To state the uses in detail is beneficial to the student, but he should recognize that it is the educational sophistication of the writer that is being revealed, not the limitations of educational measurement. As in all science, advanced instruments suggest new uses, and new uses stimulate the creation of better-designed instruments.

The central problem of all educational endeavor is learning. From the determination of national and state educational policies to the selection of

the type of broom the janitor will use in the local school, the final criterion is the facilitation of learning. In the educational sense, learning is the process of changing human behavior in socially desirable directions. In this broad interpretation, all the functions of educational measurement are concerned either directly or indirectly with the facilitation of learning.

This chapter, however, will be limited largely to the functions of measurement in facilitating learning in the classroom situation. Some attention will be given to establishing classroom situations in which measurement can function more adequately.

Since the first four chapters of this volume are concerned with the functions of measurement, it is desirable that a brief outline of the contents of these chapters be presented.

Outline of the Functions of Measurement in Education

Over-all Educational Planning

When social planning, as it relates to education, becomes more forthright and deliberate, the role of measurement assumes greater importance in the process. For example, during World War II when the maximum utilization of the nation's manpower was essential, questions were quickly asked which could be answered only through measurement. Some of these were: What are the behavior characteristics of successful combat pilots, bomber pilots, bombardiers, navigators, and the hundreds of other classifications of specialists in the essential military and civilian services? What level of aptitude is necessary for the various types of special training? What is the distribution of the various talents and combinations of talents in the general population? How can the individual with specialized talents be most quickly and reliably identified? How can manpower be most efficiently allocated to the various essential services? Experience gained in attempting to answer these questions under the stress of an emergency situation indicates that one of the most important elements in the preparation for a war emergency would be a continuing inventory of manpower during peace.

If the efficient utilization of manpower is desirable during war, it will no doubt someday be accepted as equally important in peace. Measurement will answer many of the basic questions in the process. For example, what are the behavior characteristics of successful research workers in physics, chemistry, mathematics, and the other sciences? What are the behavior characteristics of successful surgeons, lawyers, engineers, etc.? In fact, the limit of this list today is the 30,000 job titles listed in the *Dictionary of Occupational Titles*. What levels and combinations of aptitudes are es-

sential to the development of the desired types of behavior in each vocational area? What is the distribution of each talent and talent combination in the general population? How many specialists of the various types are needed? How many schools of each type will be necessary to train them?

How far social and educational planning of this type may go in the future we cannot say, but certainly more or less uncoordinated efforts in this direction have been in evidence for some time. The report of the President's Commission on Higher Education, which emphasizes the social role of higher education in our democracy; the attempt of the Office of Scientific Personnel of the National Research Council to determine the extent of potential scientific research talent in the nation; and the battery of vocational placement tests developed by the United States Employment Service constitute evidence of the trend.

From the standpoint of the individual school which is concerned with selecting students, these problems, which have been presented in a national setting, became problems of admission to, and placement within, the institution. Measurement has become an increasingly important factor in admission policies during the past thirty years.

Educational Placement

Education has two over-all functions—the *integrative* and the *differentiative*. Integrative education is, within the limits of individual aptitudes, designed to make people alike in their ideals, values, loyalties, virtues, language, and general intellectual and social adjustment. It is frequently referred to as "common education" or "general education." It unifies and gives cohesion to the social group. Differentiative education is designed to make people different in their competencies, to prepare them for the professions and specialties. The elementary school is concerned entirely with the integrative function; the high school, to a high degree; while at the college level general education tends only to supplement professional and special education.

From the standpoint of the use of measurement in educational placement this distinction is important. In the common schools, which all attend, where general education is emphasized, it is the obligation of the school to adapt the curriculum to the aptitudes and abilities of the students. In the professional schools students are selected in terms of their ability to succeed in an established curriculum.

Since World War I measurement has had an increasingly important place in the admission policies of colleges and professional schools and in the awarding of scholarships and degrees. This topic will receive detailed at-

tention in chapter 4 of this volume. The uses of measurement in adapting the curriculum to the needs of pupils in the common schools will receive attention later in the present chapter.

Guidance and Counseling

Whereas the professional schools and colleges in their admission policies are concerned with measurement from the standpoint of selecting students who will succeed in a given curriculum, the student counseling service is concerned with measurement as an aid in helping the individual student find the vocation, college curriculum, and general social environment which will make for his successful adjustment. The student's intellectual aptitudes, motor aptitudes, vocational interests, personality traits, study skills, social skills, and achievement profile are assessed with a view to assisting him to make an optimum vocational, educational, and social adjustment. This topic will receive attention in chapter 3.

Improvement of Instruction

School administrators have long recognized that one of the most effective ways of insuring that a given educational objective will be emphasized in the classroom is to measure periodically the extent of its realization. Likewise, the teacher has recognized that students will learn most effectively those things that are tested. However, it is only within recent years that the full power of measurement to modify and improve instructional procedures has been realized. The test builder has been the first to recognize the necessity for clearly formulated educational objectives described in terms of changes in behavior which can be objectively observed through student responses to test items. The tests, in turn, have clarified and emphasized the refined objectives in the thinking of teachers and students. Educational goals have become more definite and meaningful. As a result, the selection of content and the organization and nature of learning experiences have been appraised and modified in terms of the effectiveness with which specific objectives are achieved. Attention has been focused on achievement, and educational method has become a means rather than an end.

A detailed analysis of the functions of measurement in improving the goals, content, organization, supervision, and administration of instruction is presented in chapter 2.

Improvement of the Learning Situation

Measurement is a fundamental tool of educational research. As better instruments are designed and constructed, the science of education moves forward. What is known regarding the principles of human growth and

development, the nature and extent of individual and trait differences, the learning process, and the dynamics of group behavior is dependent largely upon measurement. The facts, principles, and relationships established should constantly serve as basic data in the critical examination of prevailing school organization—its objectives, procedures, and basic assumptions.

A bowing acquaintance with this fund of knowledge is certainly necessary to the practical schoolman who proposes to use measurement as a tool in increasing the effectiveness of a school. For example, imagine the possible reactions of the principal and teachers of an elementary school to the scores on a battery of achievement tests administered to the sixth-grade pupils when it is found that there is a range of eight years in reading ability in this grade and that the typical pupil varies six years in his achievement in the different subjects. Or imagine the principal who finds it necessary to set up a combined fourth and fifth grade. He combines the high achievers in the fourth grade with the low achievers in the fifth and insists that the teacher follow the prescribed course of study with both groups. At the end of the year a battery of tests is given, and the fourth-graders show superior achievement on every test. Such test results make sense only to individuals who know the elementary principles of child development and the basic facts regarding the extent of individual and trait differences.

It is impossible in a brief treatment of the functions of measurement to emphasize adequately its place in educational research and the necessity of a broad knowledge of this research for the proper interpretation and use of measurement in the school. Some of the most essential information for a rudimentary interpretation and use of test results in the practical school situation is treated in the present chapter. It is concerned with four main topics—the functions of measurement in (1) establishing learning situations appropriate to the needs, abilities, and potentialities of the individual student; (2) the diagnosis and alleviation of specific learning difficulties; (3) the motivation and directing of learning experiences; and (4) the development and maintenance of skills and abilities.

Functions of Measurement in Establishing Individual Learning Situations

The most important characteristic of a favorable learning situation is a strong ego-involved drive (purpose) on the part of the learner to acquire the various socially approved behavior patterns. To assume that such learning can be made an end in itself is perhaps one of the most frequent errors in educational thinking. Such learning is really the means of acquiring a progressive realization of social status and prestige. What the learner wants is social position with security, favorable attention, and recognition of his

special virtues, abilities, and accomplishments. He wants success in what he undertakes and a progressive broadening of significantly related new experiences. A good school or a good learning situation is one in which these fundamental cravings of the individual are satisfied through educational experiences.

The ten-year-old who gets up an hour early to practice reading a story he has volunteered to read to the class, the algebra student who puts in overtime on the really tough problems, the student who spends extra hours on a special report, the football squad at its grueling practice, and the boy who has a paper route to earn money to buy a trumpet, to take lessons for six months, to get to play in the band and wear a flashy uniform, are all examples of the social prestige factor in the learning situation. The problem then is to make learning highly satisfying in this sense. It should not be *preparation* for prestige-getting—it should be prestige-getting.

Certain elements of the learning situation which are related to the prestige goal and which may be improved through the proper use of measurement are:

1. Classification and grouping. The first consideration is that students should be classified in such a way that no embarrassment or stigma is attached. The second consideration is that the group have common learning needs. Group solidarity resulting from common goals, common understandings, common efforts, common difficulties, and common achievements should characterize the class.
2. Individual instruction with attention to specific accomplishments and deficiencies should be possible.
3. Objective measures of achievement and progress are important.
4. The individual student should be given reading materials and problems comfortable for his level of ability. Success must be emphasized. An optimistic and encouraging attitude should prevail.
5. The student should know what his specific errors, misunderstandings, and shortcomings are. The means of self-appraisal and identification of needs should be provided.
6. The instructor should have an economical method of determining the aptitudes, abilities, physical condition, social adjustment, and self-adjustment of each student.

Promotion, Grouping, and Related Problems

With minor qualifications public education in the United States is committed to twelve years of schooling for all the children of all the people. In the twelfth year as well as the first the potential unskilled laborers, truck drivers, and janitors sit beside the embryo research physicists, journalists, and surgeons. In most schools they look at the same textbooks, hear the same discussions, pursue the same educational goals, and are marked on

the same standards. At school age the dull and the brilliant look much alike, especially to the elementary school teacher who receives from thirty to fifty new pupils each year, and the teacher in the high school who meets from one hundred fifty to two hundred pupils each day.

How can educational measurement improve the effectiveness of such schools? In answering this question, it must be understood that tests are tools, the value of which depends upon the educational insight and ingenuity of the user. They are potentially as harmful when used improperly as they are valuable when used properly. There is no magic in mere use.

When a new instrument becomes available in any field of endeavor, it is natural that its usefulness be estimated in terms of its power to facilitate the achievement of the then-accepted objectives, within the prevailing forms of organization and procedure. The fact that the results of the use of the instrument constantly point to the need for revising objectives and changing organization and procedure may go unheeded for some time. The influence of measurement in and on the schools has followed this pattern. Objective tests were first used simply as examinations had always been used. But in recent years it has been recognized that, if educational measurement is to achieve its maximum value in the improvement of the educational process, the results of measurement must be considered basic data in a re-examination of prevailing school organization, objectives, procedures, and basic assumptions. Primary data for such a re-examination are those having to do with the nature and extent of individual and trait differences. A brief review of these data is essential to understanding the problem of educational grouping.

Range of intelligence by age and grade levels

One of the most carefully selected samples of representative school children available for study is the one used by Terman and Merrill (*34*) in standardizing the 1937 revision of the Stanford-Binet tests of intelligence and reported by McNemar (*25*). Seventeen communities in eleven states were sampled. Care was taken to avoid selective factors. All American-born children of the white race who were within one month of a birthday were tested regardless of grade location. Highly reliable intelligence test scores were available for this sample, since both forms (L and M) of the revised tests were administered and average scores reported.

Of the total sample, 2,106 subjects were in grades one to twelve and within the age limits six to eighteen. The median chronological age range in each grade was six years. Table 1 shows the mental age range (2nd to 98th percentile) both by chronological age and grade level for this group. One may conclude from these and other data presented in this study that

in a typical school: (1) the first-grade teacher will find that 2 percent of the pupils have mental ages of less than four years and that 2 percent will have mental ages of more than eight years; (2) the sixth-grade teacher will find that 2 percent of the pupils have mental ages of less than eight years and that 2 percent will have mental ages of more than sixteen years; (3) the high school teacher will find a range of from eight to ten years in mental age at each grade level; and (4) these conditions will be found to exist whether the school enforces strict policies of promotion and failure or promotes entirely on the basis of chronological age.

TABLE 1

MENTAL AGE RANGE BY SCHOOL GRADE AND CHRONOLOGICAL AGE*

Grade	Mental Age Range, 2nd to 98th Percentile	Age	Mental Age Range, 2nd to 98th Percentile
		18	10.0
12	8.4	17	8.0
11	8.4	16	10.4
10	7.6	15	10.8
9	8.4	14	8.8
8	8.0	13	9.2
7	7.2	12	9.2
6	6.4	11	7.6
5	5.6	10	6.4
4	5.2	9	5.6
3	4.8	8	4.8
2	4.4	7	4.0
1	3.6	6	2.8

* Adapted from Tables 7 and 8, McNemar (*25*, pp. 33–34).
The range at the extreme grade and age levels is restricted by the nature of the sample. For example, the range of 2.8 for the six-year age group is based on a sample of 130 six-year-olds in the first grade and one six-year-old in the second grade; it disregards 53 six-year-olds in the kindergarten and 22 not in school of the total sample. In schools enrolling all six-year-olds in the first grade, the 2nd to 98th percentile range would approximate four years

McNemar concludes that since both in 1916 and 1937 Terman and his associates found grade levels to vary substantially as much as age levels in mental age, "adjustments in the provisions for learning should be made without too much departure from normal grade promotions" (*25*).

Range of achievement by age and grade levels

Since the beginning of educational measurement no fact has been more frequently revealed, and its implications more commonly ignored, than the great variability in the achievement of pupils in the same grade. Anyone who has administered a battery of achievement tests to the pupils in a graded school has been struck by the great overlapping between grades. An analysis of the norms for any battery of achievement tests will reveal the extent of grade variability.

The variation in achievement based on studies by Lindquist (*22*), Cook (*5*), and Cornell (*8*) may be summarized briefly. It follows closely the

findings on intelligence. The range in achievement (2nd to 98th percentile) at the first-grade level is between three and four years; at the fourth-grade level, between five and six years; and at the sixth-grade level, between seven and eight years, in the areas of reading comprehension, vocabulary, the mechanics of English composition, literary knowledge, science, geography, and history. In arithmetic reasoning and computation the range is somewhat less, between six and seven years at the sixth-grade level. These conditions are found to prevail whether the study is based on a large standardization sample representing many schools or on the median variability of single classroom groups.

A graphic portrayal of overlapping in grade achievement of two typical eight-grade elementary schools, one with high rate of retardation, the other with low rate of retardation, is presented in Figure 1. It was developed by Cook (*5*) in a study of eighteen school systems to determine the effect of promotion policies on the mean achievement and variability of elementary school classes.

An examination of Figure 1 reveals that the chronological age distribution for the first two grades is almost identical in the two schools, but beginning with grade three there is progressively more retardation in one of the schools. The classes are most homogeneous with respect to chronological and mental age. The mental age range[1] is not surprising, but variability in achievement is so great as to be almost unbelievable to persons inexperienced in administering such tests. Least variability in achievement is found in arithmetic, with a range of six years at the fourth-grade level extending to eight years at the eighth-grade level, but in English, reading, science, geography, and history almost the complete range of achievement is found in each grade above the primary. The mean achievement for the school with the low retardation rate is significantly higher in most subjects than in the school that retains the low-ability pupils longer. The range of achievement with which each teacher must cope is not significantly different in the two schools. The most important single generalization that may be drawn from a study of this chart is that in the primary grades a teacher may expect a range of from four to five years in achievement, while above the primary level almost the complete range of elementary school achievement is present in every grade.

It should also be noted that the low achievers in any four or five suc-

[1] Mental age was measured with the Kuhlman-Anderson Intelligence Test. This test tends to yield a narrower range of mental ability at age and grade levels than other intelligence tests, probably because only 10 tests of a possible 39 in the entire scale are given to any group. The test manual cautions against this practice, but not with sufficient emphasis. The use of the median mental age on the 10 tests instead of the mean may also tend to reduce the variability. Achievement was measured with the Unit Scales of Attainment battery.

cessive grade groups have more in common than they have with the mean achiever of their own grade. Hence, if the low achievers in the eighth grade were demoted to the fourth, they would still be low achievers in the fourth grade. The same is true regarding high achievers in any several successive grades. They resemble each other more in achievement than they do the mean pupil of their grade.

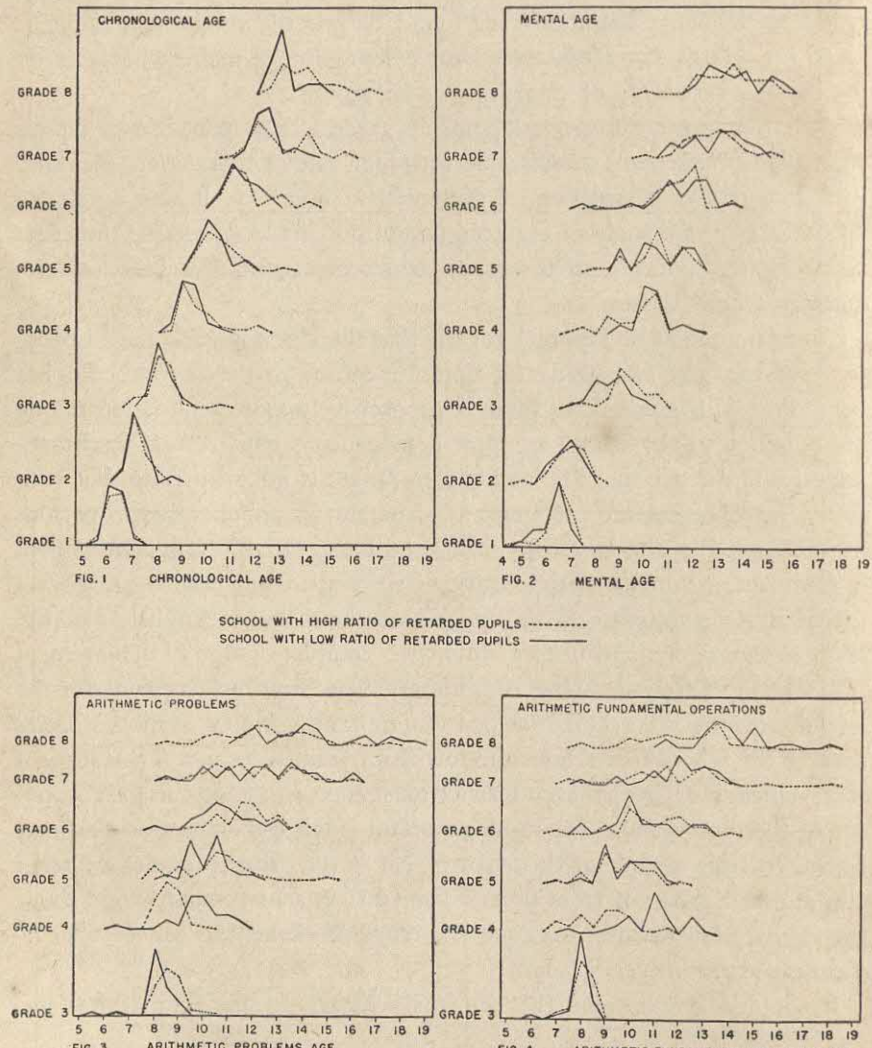


FIG. 1.—Variability in chronological, mental, and achievement ages of pupils in two eight-grade elementary schools, one with high rate of retardation, the other with low rate of retardation. Frequencies computed as percent of class at each age level. (*From* Cook [5, pp. 28–29].)

The policy of retardation and acceleration, which still prevails in many schools, is based largely on the assumption that the practice results in grade groups which are more homogeneous in achievement than age groups

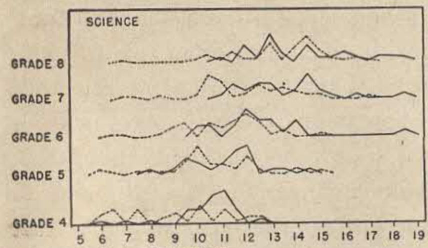


FIG. 5 SCIENCE AGE

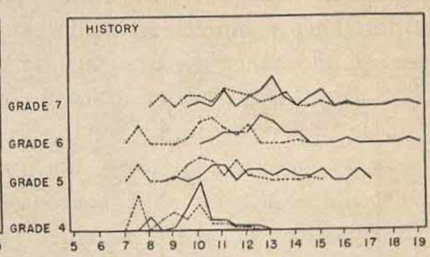


FIG. 6 HISTORY AGE

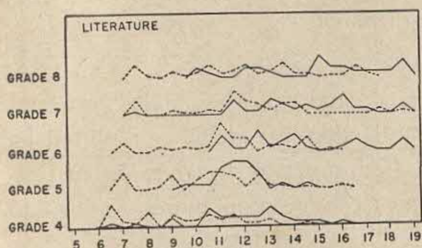


FIG. 7 LITERATURE AGE

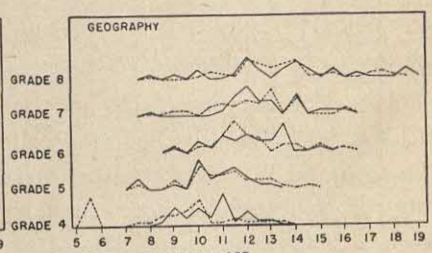


FIG. 8 GEOGRAPHY AGE

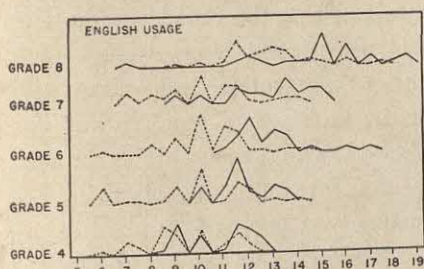


FIG. 9 ENGLISH USAGE AGE

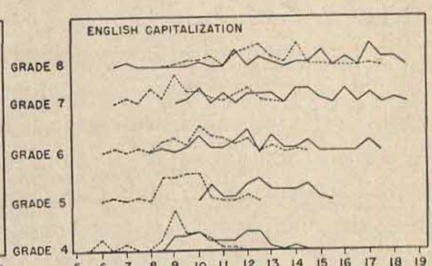


FIG. 10 ENGLISH CAPITALIZATION AGE

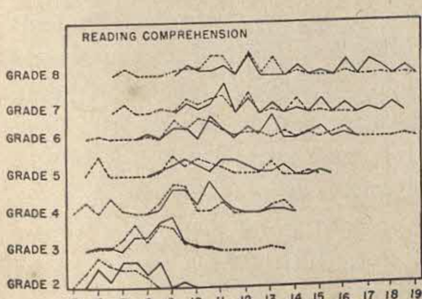


FIG. 11 READING COMPREHENSION AGE

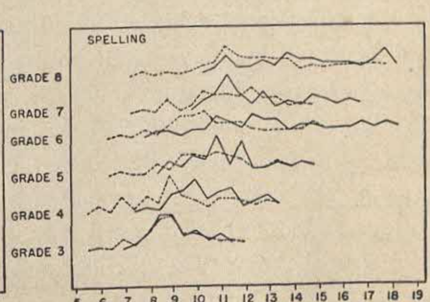


FIG. 12 SPELLING AGE

FIG. 1.—*Continued*

would be. Although there are few reports of the variability in achievement of unselected age groups, what evidence is available does not support this assumption.

Cornell (*8*) reports the variability of New York State school children in intelligence and achievement at three age levels—seven, ten, and fourteen. After comparing these age data with grade data from two previous studies by Coxe (*9, 10*), Cornell concludes that, "both the range of the middle 50 percent and the total range show no marked differences in favor of either age or grade groups. For practical purposes of classification, then, we could deal with an age group without any more difficulty due to diversity than we find in a grade" (*8*). This conclusion is similar to that reached by McNemar in comparing the age and grade ranges in intelligence.

Variability of high school and college students in intelligence and achievement

The most comprehensive and detailed study of the intelligence and achievement of high school and college students available is reported by Learned and Wood (*19*). It involved from eight to twelve hours of testing for 26,548 high school seniors, 5,834 college sophomores, and 3,826 college seniors, all in the state of Pennsylvania.

Intelligence was measured by the Otis Self-Administering Test of Mental Ability. A study of the findings (*19*) reveals that the broad range of intelligence found in the lower schools is also present in high school and college. The total range of intelligence at the college sophomore and senior levels was almost as great as at the high school senior level. The college distributions were, of course, skewed, with a much higher proportion of students at the higher levels. With reference to the college sophomore distribution, the college senior median was one-half standard deviation above the sophomore median, and the high school senior median was approximately three-fourths of a standard deviation below it.

The educational achievement of the high school and college students was measured by the General Culture battery, consisting of tests in general science, foreign literature, fine arts, and social studies. The overlapping of high school and college classes in achievement as measured by these tests is most concisely revealed by the following facts: More than 28 percent of the college seniors had achievement scores below the average college sophomore, and nearly 10 percent had scores below the average high school senior. The distribution of scores for the high school seniors revealed that 22 percent of them exceeded the average college sophomore and that 10 percent exceeded the college senior average.

In order to measure the variability in achievement of classes within the same institution, one representative college was studied intensively. A battery of tests was administered to the entire undergraduate student membership. The battery consisted of the General Culture Tests already mentioned plus a two-hour test in English literature, vocabulary, and usage, and a test of two hours and twenty minutes in mathematics. It was found that if the graduating class that year had been selected from the entire student body on the basis of achievement test scores instead of from the senior class on the basis of credits earned, only 28 percent of the senior class would have graduated. The remainder of the graduates would have been made up of 21 percent of the juniors, 19 percent of the sophomores, and 15 percent of the freshmen. The mean score of the graduating class selected on the basis of achievement would have been one standard deviation above the average of the class that actually graduated, and its mean age would have been two years younger.

The range of general culture achievement with which a college instructor must cope in a typical class is indicated by the fact that in 50 percent of the college sophomore classes, the range of achievement extended from a point below the 25th percentile of high school seniors to a point above the 75th percentile of college seniors.

The Problem of Individual Differences

Emphasis has been placed on the variability of age and grade groups in intelligence and achievement because, in spite of the overwhelming evidence available, educational practice, if not educational thinking, has in the main tended to ignore it. The basic idea persists that grade level as determined by time spent, exercises performed, and courses passed is closely related to intellectual skills, understandings, and usable information.

Many of the false assumptions which inhibit teachers in meeting the needs of individual pupils are corollaries of the idea that grade levels signify rather definite stages of achievement. Despite the fact that in the elementary school the typical range of intelligence and achievement in a grade is from four to eight or more years, the teacher is commonly assumed to be a grade specialist with specific knowledge and techniques appropriate to a given grade and, hence, should not be expected to teach pupils who deviate markedly from that level. It is, furthermore, assumed that the course of study for a grade is the scheduled academic requirement to be administered uniformly to all pupils in the grade, that all pupils should be capable of coping successfully with the work outlined for the grade, that a pupil should not be promoted to a grade until he is able to

do the work of that grade, and that when individual differences are provided for, all pupils can be brought up to standard. It is still believed by many that the strict maintenance of the passing mark results in relatively homogeneous classes, that satisfactory instruction for a class can be based on uniform textbooks, and that when relative homogeneity does not characterize a class, it is the result of poor teaching or lax standards. Since schools were first graded, such assumptions have thwarted teachers in the process of adapting instruction to the abilities of individual pupils.

Despite the fact that the teacher is supposed to understand individual pupils—their aptitudes, abilities, deficiencies, interests, and peculiarities—it is not uncommon practice to assign an elementary teacher from thirty to fifty new pupils each year (twice a year where semiannual promotion is still practiced). If grade specialization were considered less important and knowledge of pupils more important, the teacher would remain with the same pupils for several years. At the high school level, where variability in achievement approaches its maximum, the teacher really becomes a departmentalized specialist in subject matter. Here the teacher with the special knowledge, techniques, and textbook meets from one hundred and fifty to two hundred different pupils each day. The struggle here is no longer to know the abilities, aptitudes, and interests of pupils but only their names. In college the attempt to know names is too often given up.

Attempts to solve the problem of individual differences in the schools take two general forms. The first, which has received major emphasis in the past, assumes that instructional groups can be made relatively homogeneous with respect to general ability and then subjected to standardized educational methods using uniform textbooks, assignments, recitations, and examinations adjusted to the level of achievement. The procedures relied upon for forming such groups are: (1) grouping on the basis of one or more measures of general academic aptitude, such as intelligence tests, general educational achievement tests, school marks, and teacher opinion; (2) judicious policies of acceleration, promotion, and failure; and (3) effective teaching.

The second approach assumes that heterogeneity resulting from both trait differences within the individual and variation between individuals is so great that traditional mass instructional procedures must be discarded in favor of techniques designed to meet individual needs. The goal here is to know and accept the great variability of instructional groups as it exists, under even the best of grouping procedures, and then to discover effective methods of providing for individual needs and capacities in such heterogeneous groups. Measurement has an important role in the process.

Effectiveness of general-ability grouping

American educational procedure for the past century has been largely concerned with contriving methods of uniform instruction for postulated homogeneous groups of pupils. The period from 1800 to 1850 was one of intense educational activity and reform. New states, from Ohio to Iowa, were adopting constitutions with educational provisions; all were taking their educational responsibilities seriously; some were sending experts to Europe to study schools. Horace Mann was at work in Massachusetts, Henry Barnard in Connecticut. Some of the results of this activity were the graded eight-year elementary school, grade teachers, graded textbooks, school buildings designed to house graded programs, normal schools established to teach the techniques of mass instruction, and the creation of the positions of state, county, and city superintendent of schools to direct and coordinate the organization. By 1870 the schools of the United States, even the one-room rural schools, were graded, and uniform state courses of study were in vogue. Achievement standards for each grade were determined subjectively by the authors of the graded texts and courses of study. Promotion standards were strict, and before long "laggards in our schools" became a national concern.

Numerous remedies for uneven educational progress were tried between 1875 and 1925. Some of the earlier remedies had to do with promotional policies. Semiannual promotion, quarterly promotion, subject promotion, and special promotion each had a trial. It was found that the less severe the effects of nonpromotion were made, the more it was practiced. Some remedies attempted to hold standards constant and increase the amount of instruction for slow pupils, as in the Batavia plan, the assisting-teacher plan, and the vacation-classes plan. In some, the course of study was held constant, and the amount of time required for slow-, medium-, and fast-learning pupils was differentiated, as in the North Denver plan, the Cambridge plan, and the Portland plan. In others, time-in-school was held constant, and the course of study differentiated for slow-, medium-, and fast-learning pupils, as in the Santa Barbara and Baltimore plans. Other plans involved dividing the course of study in each skill subject into units of specified activities and achievements, permitting each pupil to advance at his own rate in each skill, with group instruction in content areas, as in the Pueblo plan, the Winnetka plan, and the Dalton plan. Several of these plans require that pupils be grouped into slow-, medium-, and fast-learning groups. The ability to do this seems not to have been seriously questioned at the time.

After World War I the rapid development of group intelligence and

achievement tests stimulated the practice of general-ability grouping in the elementary and high schools. The plan was first used on a large scale in the schools of Detroit in 1920. In that year the 10,000 pupils entering the first grade were sectioned on the basis of scores on a group intelligence test into three groups designated as X, Y, and Z. The upper 20 percent of the children were designated as the X group, the middle 60 percent as the Y group, and the lower 20 percent as the Z group. Differentiated curriculums were provided for each group with the aim of securing the best possible experience for the entire range of ability represented.

The idea of grouping pupils on the basis of some measure of general ability was extensively adopted. The degree to which curriculums were differentiated to meet the needs of the hypothetical slow, average, and bright pupil varied from system to system. Bases for grouping also differed, as did the proportion of pupils placed in the different groups. In 1936 the National Society for the Study of Education devoted Part I of its *Thirty-Fifth Yearbook* to a critical evaluation of practices in the grouping of pupils. General-ability grouping was criticized and defended on educational, philosophical, social, and psychological grounds. Experimental studies were summarized and the conclusion reached that the evidence slightly favored ability grouping where adaptations of standards, materials, and methods were made. The pupils in dull groups seemed to profit slightly, with pupils in bright groups frequently showing lower achievement than when placed in more heterogeneous groups. However, little attention was given to the question of the extent to which general-ability grouping reduced the variability of instructional groups in specific subject areas.

General-ability grouping is based on the hypothesis that there is relatively little variation from trait to trait within the individual, that all traits with which the school is concerned are substantially correlated, and that mental functions are organized around a predominating general factor which determines the general-competence level of the individual. Evidence from several overlapping fields of investigation tends to refute this hypothesis. One is concerned with basic theories of mental organization, a second, with studies of so-called *idiots savants,* a third, with asymmetry of development in "normal" and "gifted" individuals, a fourth, with direct measures of trait variability, and a fifth, with evidence of correlations between traits. This research has been well summarized by Anastasi and Foley (*2*). General conclusions for our purposes may be based on Hull's (*15*) study of the variability in amount of different traits possessed by the individual. His subjects were 107 ninth-grade boys to whom he adminis-

tered thirty-five standardized psychological and educational tests involving a wide variety of functions. Hull concludes that trait differences in the typical individual in this group were 80 percent as great as individual differences in the total group, that trait differences tend to conform to the normal curve, are more than twice as great in some individuals as in others, and that no relationship exists between the individual's general level of ability and the extent of his trait variability.

Studies of the extent to which the variability of elementary school classes can be reduced through general-ability grouping have been made by Hollingshead (*14*) and Burr (*4*). Hollingshead was primarily concerned with determining the best measure of general ability for classification purposes and Burr with the extent to which general-ability grouping reduces variability in reading and arithmetic. Educational achievement test batteries were found to be the most effective basis for grouping. The variability of the typical X, Y, and Z section in reading and arithmetic was found to be approximately 80 percent of the total grade range. Individual pupils were found to be such complexes of more or less independent abilities that when sections were made nonoverlapping in one phase of a subject such as arithmetic reasoning, they overlapped greatly in another phase such as arithmetic computation.

The important generalization to be drawn from studies of trait variability is that instructional groups formed by general-ability grouping are not sufficiently homogeneous to warrant uniform mass instructional procedures. For example, a typical sixth-grade class has a range of eight years in reading ability. After X, Y, Z sectioning on the basis of educational age, each section will still show a range of from five to seven years. Regardless of grouping procedure the teacher's attention must constantly be directed to individual children and their immediate problems in learning. The obligation of the school to furnish instructional materials with a range of difficulty commensurate with the range of ability in a class and to individualize instruction is as great when general-ability grouping is practiced as when it is not.

Effectiveness of judicious policies of promotion

Teachers and administrators at the high school and college levels frequently attribute the wide range of ability which they find in classes to lax promotion practices in the lower schools. They maintain that the practice of universal promotion which has crept into the lower school to hide its failures and keep the community happy is responsible for the wide range of individual differences in the upper school. What is needed, they argue,

is meaningful grade standards which must be achieved before a pupil is promoted. This would be an honest policy, they argue, and would require pupils to earn their promotions. Grade levels would mean something and really signify definite stages of educational development. More rigorous policies would raise the standards of our schools. Pupils would be stimulated to greater effort. According to this line of argument, pupils who are promoted regardless of achievement suffer emotionally from a progressive inadequacy to deal successfully with the school situation. They become discouraged, quit trying, and learn that they can "get by" without effort. If individual differences are provided for, they conclude, almost all pupils can attain a respectable standard. We must provide for those differences and get all pupils up to that standard. An investigation reported by Cook (7) tests the validity of several of these claims. Complete test records were available for 148 Minnesota school systems. These systems were first ranked on the basis of the amount of retardation. Then nine systems that approached the universal promotion end of the scale were matched with nine systems which maintained rigorous standards of promotion. Matching was on the basis of size of school, socioeconomic status of parents, and professional qualifications of teachers.

It was found that schools which have relatively high standards of promotion (retard the dull and accelerate the bright) tend to have a higher proportion of over-age, slow-learning pupils, since such pupils remain in school from one to several years longer. The higher proportion of such pupils reduced significantly the mean mental age and achievement level of grade groups in these schools.

In comparing the variability of the two groups of schools on eleven achievement tests and intelligence at the seventh-grade level, no significant difference was found. The higher proportion of low-ability pupils in the schools with high rates of retardation tended to keep the variability of classes large. Pupils were rarely failed more than twice in the same grade and eventually reached the upper grades in spite of the efforts to maintain standards.

When the achievement of pupils of the same age and intelligence in the two groups of schools was compared, there was no reliable difference. This would tend to indicate that the schools were well matched and also that the constant threat of failure did not increase achievement in the schools attempting to maintain high standards. It also emphasizes the fact that the retention of a large number of low-ability pupils through nonpromotion tends actually to reduce grade standards and aggravate the range of ability problem.

As long as all the children of all the people remain in school, it will be impossible to reduce the variability of instructional groups significantly through promotion policies. If rigorous policies of promotion are adhered to, the efficiency of the school is reduced through the accumulation of low-ability pupils and the lessening of educational opportunities for the more able.

Effective teaching procedures influence individual differences

The idea that instructional groups can be made more homogeneous in a given achievement area through effective teaching is quite common. It keeps company with other ideas inherent in the traditional conception of the schooling process. According to this point of view the course of study outlines what is to be taught in each subject each year, what textbooks are to be used, and what pages are to be covered by a given date. Education consists of learning such things as are found in courses of study and textbooks: spelling words, type problems in arithmetic, names, dates, causes and results of wars, states and capitals, explorers and where they explored, cities and their characteristics, countries and their products, rules for punctuation and capitalization, and the seven basic food groups.

Conscientious teachers with this point of view who have taught a given grade for a number of years know this material well, and they know how to threaten, coax, drill, drive, review, and test pupils until they learn it. There is a passing mark, perhaps 70 or 75, and this is interpreted by the teacher as meaning that this percentage of what is taught must be learned in order to pass to the next grade. The aim is to get all pupils over the passing mark, to make them homogeneous with respect to the goals, to provide for individual differences in the class, and to get them all up to standard. The conscientious teacher does these things. Why, then, should the range of achievement be what it is in these classes? What is wrong with these goals? Is there harm in striving for uniformity of achievement? Can homogeneous instructional groups be achieved in this way?

The problem with which we are dealing is basically concerned with the effect of a period of learning upon individual differences. Are individuals more alike or less alike with respect to a given ability after a period of instruction? Does good teaching increase or decrease the variability of a class? A considerable amount of research on this question is available and has been summarized by Anastasi (*1*), Peterson and Barlow (*30*), and Reed (*31*). The research indications are somewhat contradictory, and many technical problems are involved in their interpretation. But for our purposes the following generalization seems warranted: if the responses

to be learned are sufficiently simple and the goals that have been set so limited that a high proportion of the group can master them during the period of learning, the variability of the group becomes less; if the task is complex and the goals unlimited, so that the abilities of the most apt members of the group are taxed during the period of learning, the variability of the group increases.

The point of view that instructional groups can be made more homogeneous in a given achievement area through effective teaching presupposes limited educational objectives—limited not only with respect to the complexity of the learning but also with respect to the amount of even the simple learnings. In order to place requirements within the reach of the less able student, emphasis is commonly placed on rote factual learning and type problem solving. Textbooks in the content areas are frequently written down to a difficulty level of two or three years below the grade at which they are to be used. Materials are selected for inclusion in each grade curriculum which can be mastered by at least 80 percent of the pupils. Questions and answers are drilled upon before examinations with the understanding that the final examination questions will be taken from the list. Frequently the teacher's time is devoted almost exclusively to the slow learners. In such schools when pupil progress is measured over a period of time in terms of the limited goals, pupils with the lowest initial scores make the greatest progress. There tends to be a negative relationship between initial scores and gain. The relationship between gain and intelligence is low and frequently negative. These tendencies indicate that pupils of high and even average ability are understimulated and essentially neglected.

The most serious result of emphasizing limited goals in education is that what is learned is frequently of little value and is retained for but a short period. Only rote memory, a low-order mental process, is required to pass the *name, describe, define, who, what, when,* and *where* type of examination with which achievement is frequently measured. Tests of retention given from three months to three years after a course is completed reveal that from 40 to 80 percent of the information required by the final examination is lost. Buckingham (*3*) has called this the greatest waste in education. Investigations in algebra, botany, chemistry, physiology, psychology, and zoology all reveal the same rapid rate of forgetting. The forgetting curves closely approximate those for nonsense materials, indicating that much of what is learned for examination purposes is no better organized and no more useful than nonsense materials.

The relative permanency of different types of learning has been investi-

gated by Tyler (*38*) and Wert (*39*). In Tyler's study a test in zoology measuring five objectives was administered to eighty-two students at the beginning of the course, at the end of the course, and fifteen months later. Percent of loss or gain on each part of the test during the fifteen-month period was computed in relation to the amount gained during the course. On the part of the test requiring: (1) names of organs identified from pictures, the loss was 22 percent; (2) recognition of technical terms, the loss was 72 percent; (3) recall of facts, the loss was 80 percent; (4) application of principles, there was no loss or gain; (5) interpretation of new experiments, there was a gain of 126 percent. Wert's experiment, also in zoology, is quite similar. This experiment measured percent of loss or gain over a period of three years in relation to the amount gained during the course. A gain of 60 percent was found in application of principles to new situations, and a gain of 20 percent in interpretation of new experiments. There was a loss of over 50 percent in terminology, function of structures, and main ideas; and a loss of over 80 percent in associating names with structures.

These experiments indicate that learning involving problem-solving relationships and the operation of the higher mental processes are relatively permanent and that unrelated facts and mere information are relatively temporary. Unless learning involves differentiation and integration of old and new responses into a problem-solving type of mental process or into an organized behavior pattern, it has little permanence or value. *How was it learned?* is the important question.

The permanent results of the educational process are measured by such tests as: (1) vocabulary, (2) reading comprehension in the natural sciences, social sciences, and literature, (3) problem-solving in mathematics and the sciences, (4) ability to use the library and basic reference materials, and (5) ability to write and speak effectively. It is with reference to such objectives that great heterogeneity in achievement exists, and the more effective the instruction, the greater the heterogeneity.

It would seem then that the emphasis which some schools place on striving for homogeneity in classes, getting students over the passing mark, and providing for individual needs with a view to bringing all pupils up to a standard, encourages teachers to set limited goals for instruction which result in temporary factual learning involving mainly the lower mental processes. When the ultimate goals of education, involving the higher mental processes and permanent learnings are striven for and each student is stimulated to capacity effort, the variability of instructional groups increases.

Measurement as an Aid in Meeting Individual Needs in Heterogeneous Groups

To direct the educational experiences of each child in such a way as to bring about his optimum development and adjustment to his culture presents one of the most complex problems imaginable. Materials presented up to this point were selected with a view to revealing the extent and nature of individual and trait differences and the general effect of educational experiences upon them. Since many of the educational devices and procedures directed to this end have in the past sprung from misconceptions regarding the facts and have oversimplified the problem, this approach seemed necessary. Recommendations for the uses of measurement in solving the problem must begin with these facts and be limited by them. The approach must be experimental and tentative to a high degree, since interpretations and implications, even when drawn from facts, are subject to error.

In general social intercourse during the years of maturation, the traits most important to the child in determining membership and status in a congenial group are chronological age and general physical development. Together with general social development these are also the most obvious. In a graded school it is tremendously important to a child that he be grouped with his peers. To deny him this privilege is to violate one of the most important requirements of a favorable learning situation. Therefore, throughout the period of maturation, which corresponds roughly with the compulsory school age, these traits should constitute the fundamental basis for educational grouping, that is, when a child is five he enters kindergarten, when six he enters the elementary school, when twelve he enters the junior high, and when fifteen the senior high school. Since chronological age is not perfectly related to physical and social development, some adjustments have to be made, especially at the primary level with developmental level taking precedence over chronological age. However, the all-important factors in the basic grouping of pupils are physical and social development. A child should live and work in the group he most obviously belongs with, which accepts him and which he accepts. It will be recalled that *both intelligence and achievement test data show age groups no more variable than grade groups. Hence this grouping will not materially increase the range of ability with which the teacher must cope.*

The secondary basis for grouping should be achievement in specific areas, that is, for reading instruction children should be grouped in terms of reading ability and needs, for English instruction in terms of needs in English. It should be remembered, however, that even when grouped in

this manner attention must be focused on the individual pupil, since two pupils making exactly the same score on a test will have different abilities and needs.

If physical and social development is accepted as the primary basis for grouping in the common school, all assumptions that a grade level indicates anything specific regarding intellectual competence or educational achievement must be given up. Evidence has been presented indicating that grade level, per se, never truly signifies these things. The assumption that it does has led to absurd practices, subterfuge, and confusion in meeting individual needs. The determination of intellectual competence and educational achievement must rest primarily upon measurement. Teacher observation and judgment will always be of extreme importance in the learning situation and in the appraisal of those traits not yet measurable. But measurement will carry the burden of information on educational status.

Diplomas in the common schools will continue to be given by virtue of years attended, age attained, and courses taken, but they will be assumed to convey little or no further meaning. The common school period probably should be the same length for all pupils. No attempt should be made to bring the slow pupil up to standard by keeping him in school a few extra years. No attempt should be made to keep the bright pupil down to standard by accelerating him through school at an early age. If instruction is to be adapted to individual needs and capacities, the diploma should be considered little more than a certificate of attendance, and the level of achievement attained in the various areas should be determined through measurement procedures.

To set respectable standards of educational achievement for graduation in schools attended by all the children of all the people means automatic labeling as failures, with moral and social approval implications, of a large percentage of pupils. It also means that pupils below this level in competence will, with the encouragement of teachers and parents, attempt learnings that are beyond their capacity. The status of pupils well above the standards will not be adequately represented by the diploma, nor will their capacities be tested by the curriculum. Measurement makes possible the elimination of a set standard for all pupils who finish the common schools. The expected achievement of each pupil can be related directly to his capacity and to the requirements of his probable adult vocational position in each of the learning areas. His educational status and growth in these areas can be more accurately determined and communicated through measurement.

The general limitations of diplomas, and to some extent school marks, and percentile rank in class in accurately representing academic capacity and status for transfer and certification purposes are well known. There are several reasons for this. Teachers differ in grading standards, classes differ in quality of students, and schools differ in both grading standards and academic capacity of students. In state-wide high school testing programs it is frequently found that schools differ so much in achievement in the various subjects that the distributions of scores for the extreme schools are nonoverlapping. In ninth-grade algebra, for example, the lowest-scoring pupil in the highest-scoring school is above the highest-scoring pupil in the lowest-scoring school. The pupil who receives an F in the highest-scoring school would receive an A if enrolled in the lowest-scoring school.

Wood (*40*) has proposed a dual system of marks to meet the two related, but independent, purposes for which grades are given. He recommends that for certification and transfer purposes, comparable educational measures be based on the best standardized educational tests. And for the moral, social, and disciplinary welfare of the individual pupil he recommends that the local school marks be retained.

If local school marks are based on relative standing in a class with a wide range of capacity, their moral, social, and disciplinary value is questionable. If, however, marks are based on achievement in relation to well-established capacity, or on the achievement of objectives peculiar to the school, or those not yet measurable in an objective sense, they have an important function.

With reference to valid educational goals for which standard scales of development are available, there would seem to be little reason for indicating progress in other than objective terms. The practice in some schools of converting standardized achievement test scores into the more or less ambiguous school mark is indefensible. For example, if a pupil has made a gain in reading age of from ten years in September to ten years and nine months in January, this fact should be known to all persons concerned, including the pupil and his parents. To convert such a fact into a school mark is absurd. (Development in the measurable educational achievement areas should be considered as objectively as is growth in height and weight. Certainly the suggestion that a pupil's growth in height be compared with that of others in the same class and a grade assigned would be considered absurd.) A gain of three months by another pupil during this same period may be properly considered a greater achievement relative to capacity. Individual development becomes the all-important consideration. When a testing program makes such measures of achievement sys-

tematically and periodically available, traditional school marks lose much of their significance. Also the types of educational experiences which produce the greatest development are more quickly and accurately determined.

The use of objective measurement fosters the child-development point of view and makes obvious the necessity of adjustment of instruction to individual capacity. Such a point of view tends not to develop in a school which assesses educational progress in terms of subjective marks based on relative standing in a class.

The emphasis placed on the value of measurement in guiding educational development in the schools attended by all children leaves no place for the use of measurement as a flunking device or as a basis for exclusion from school except at levels so low that institutional commitment is indicated. In professional schools and colleges measurement has a definite and justifiable function as an exclusion and flunking device. Individuals without the necessary aptitudes and abilities who aspire to teach school, or practice law, engineering, dentistry, medicine, and so forth, must be eliminated in the interest of the public welfare and economy.

Suggested administrative policies

The traditional graded elementary school and the departmentalized high school, with their emphasis on required curriculums,uniform courses of study, uniform textbooks, heavy pupil-teacher load and the like, were designed without regard for individual and trait differences. The purpose was to give a "shotgun" type of education to as many pupils as possible with the least expenditure of money. Educational policy was dictated largely by available funds, not by pupil needs. Even in such schools measurement has increased educational efficiency by making the teacher aware of individual aptitudes, abilities, and specific needs with an economy of time and effort. But assuming that the teacher has the "developmental" rather than the "subject-matter-to-be-covered" point of view and knows how to stimulate pupils to maximum educational effort, she still lacks classroom space, time, equipment, books, supplies, freedom, encouragement, and energy to do the job. Many highly competent and conscientious teachers leave the profession yearly because of the resulting frustration. The relatively simple problem of providing a classroom with reading material which has a difficulty range and interest appeal commensurate with the abilities and interests of the class is rarely attacked seriously.

The administrative policies designed to enable teachers to meet the needs of individual pupils in heterogeneous groups have two purposes:

(1) to make it possible for the teacher to know the pupil better—to know his abilities, his interests, and his deficiencies well enough to direct his learning; and (2) to provide instructional materials with a range of difficulty and content commensurate with the range of abilities and interests of the instructional group.

1. The testing program of the school should be systematic and comprehensive. It should furnish the teacher with up-to-date information regarding the growth record and status of each of his pupils in at least the fields of English, reading comprehension, mathematics, study skills, and problem-solving in the natural and social sciences. The tests should measure at regular intervals the permanent learnings which have been achieved toward the major and ultimate objectives of education. Knowledge of the pupil and his record of achievement should be considered basic data in the educational process.

Measurement should begin with a relatively undifferentiated test of the individual Binet type at the preprimary level and reach a considerable degree of differentiation in terms of improvable skills, abilities, and attitudes by the junior high school level. During the primary school years the differentiation should be in terms of specific abilities related to developing number concepts, reading readiness, some aspects of beginning reading achievement, and specific behavior related to health and socialization. Group testing which assumes reading ability is generally unsatisfactory below the fourth grade. During the intermediate school years, however, the approach to complete differentiation in terms of ultimate educational objectives can be made. The selection of tests should emphasize the more permanent learnings and skills and discourage mere factual learning.

The results of the testing should always be portrayed in graphic profile form, showing the growth in each differentiated ability from year to year. This test record should be in the hands of the teacher in the permanent record folder, which should contain information on other important aspects of the child's development, the health record, and information on social and emotional development. It should also include pertinent information about the child's family, his interests, and activities outside of school, as well as samples of the best art, handwriting, poetry, and composition produced each year.

Systematic testing of the type described above should preferably be done at the beginning of the school year. This time of testing has several distinct advantages. It provides the teacher with up-to-date information regarding the educational status of each child at the beginning of the period of instruction when attention should be focused on planning in terms of indi-

vidual needs. The temptation to use the tests to determine promotion or failure is minimized. When a child knows his status in improvable skills at the beginning of a year and considers his progress during the preceding year, there is usually an urge to better his previous record. When tests are given at the close of the school year the efforts of the teacher are more likely to be centered on preparing children for the tests and result in undesirable "cramming" procedures.

Other tests of a more diagnostic nature in all the basic learning areas should be available to teachers at all times in order to determine individual needs and measure progress in the skills and abilities being emphasized in instruction. These tests should always be selected and used strictly from the standpoint of their instructional value.

2. Grouping within the class on the basis of status in specific learning areas is one of the most essential procedures in meeting individual needs. In the elementary school from three to five groups within each class are common, with the pupils grouped differently in each subject or ability area. The within-class grouping in reading, for example, may be based on ability in general reading comprehension or on a more detailed analysis of individual needs, or both. Some of the groupings may be continued over a relatively long period; others may be for only a day or two or until the need for them no longer exists. Pupils may be transferred in or out of a group at any time. Grouping should be practiced not only in terms of needs in skill areas but also in terms of interests in a topic, or in terms of personality and social needs. The latter groups commonly take the form of committees selected to carry through a special project. Individual assignments of responsibility have the same purposes. Much of the work of the class, however, will still involve the entire group. This is necessary in planning, coordinating, and unifying the over-all activities and goals of the class. The unification of a class is achieved largely through the unit activities. The pupil needs to feel most strongly his membership in, and acceptance by, the total group, the other groupings being for special needs and purposes which he considers subsidiary to the over-all purposes of the total group.

3. The teaching load must be reduced. Teachers skilled in the art of adapting group procedures to the needs of individual pupils insist that no primary teacher should be responsible for more than twenty-five pupils and that thirty pupils should be the maximum load of intermediate grade teachers. At the high school and college levels the teacher-student load determines how much individual attention is possible.

4. Many elementary schools permit the teacher to remain with the same

pupils for more than one year. In some schools the pupils have the same teacher from kindergarten through the sixth year; in others, one teacher through the primary years, and another through the intermediate years. The continuing-teacher plan eliminates the concept of a grade teacher and places emphasis on knowing the pupil and his needs, knowing parents, and thinking in terms of child development. The teacher is enabled to start each year with a thorough knowledge of his pupils and can plan the work of the year in terms of specific needs. The process of promotion is eliminated.

At the junior and senior high school levels where the typical teacher meets 150 pupils each day, the problem of making it possible for teachers to know their pupils better can be alleviated to some extent by having home-room teachers teach their home-room groups for at least two periods, preferably in the core subjects—social studies and English. These two subjects should be integrated, utilizing the social studies and literature as content and stimulation material in learning to read, write, speak, and listen effectively.

5. The plan of reporting to parents percentage marks, or letter grades, is not consistent with the policy of meeting the needs of individual pupils. Likewise, the practice of simply marking a pupil satisfactory or unsatisfactory in terms of his general learning capacity is inadequate. The weakness of these marking and reporting methods is not that they tell too much about the pupil, but rather that they tell too little and their meaning is ambiguous.

It is more beneficial and revealing to all concerned if the teacher once or twice a year, oftener when desirable, sits down with the parents and discusses the child's achievements and needs, going through his record folder and noting measures of growth and status, his achievement in the basic skills and toward the major educational objectives, discussing samples of his work, his behavior characteristics, and his personality needs.

6. A wealth of instructional material must be provided. Instructional materials must have a range of difficulty, interest appeal, and content commensurate with the range of abilities and interests of the class. In the elementary school, classroom libraries must be provided which contain the basic reference materials, books at appropriate levels of difficulty on the units to be developed by the class, children's literature in abundance, and books on popular mechanics, hobbies, and how to make things. The school library should constantly feed into and supplement the room libraries, but it should not supplant them.

In addition to a wealth of books, there must be in the elementary classroom children's magazines and newspapers, art materials, tools, and work-

benches, simple science laboratory equipment, a science cabinet and sink, large bulletin boards, as well as blackboards, pictures, visual aids of all types, movable desks, tables and chairs, heavy and light screens for dividing the room, and a combination radio-phonograph. In the high school and college the equipment of rooms will be more specialized, but the pattern of adequacy suggested for the elementary school should be followed.

Adjustment of curriculum policies and practices

It is obvious that if in the elementary school all pupils are required to read the same books, do the same exercises, solve the same problems, and attain the same minimum standards on examinations, there can be but slight recognition of individual interests and abilities. Likewise, in the high school and college if curriculum requirements are rigid, if students are required to take certain subjects, or a sequence of courses in a given curriculum, recognition of individual needs is thwarted. This is true, whether the requirements are imposed by the state, the school-standardizing agencies, or the local school administration.

Whenever a school purports to accept all the children of all the people, it must strive for a curriculum sufficiently flexible and broadened to recognize and reward the great variety of combinations of aptitudes and interests of its pupils, enabling them to know their peculiar strengths and weaknesses and preparing them to fit into our complex society with its multiplicity of demands. The curriculum must be modified to provide flexibility of requirements, and the teachers and counselors must be free to plan for the welfare and optimum development of individual pupils.

1. In general education the curriculum content should be organized for the teachers' guidance around significant aspects of the social and natural environment for the purpose of developing understanding and intelligent behavior with respect to them. It must be recognized that the understanding of the immediate cultural and physical environment requires an understanding of other cultures in other environments as determined by evolutionary, geographical, and biological factors.

As far as pupils and their learning are concerned, the foregoing should be the result, not the starting point. For pupils, the curriculum is organized around their purposes and understandings. The teacher must guide the pupils' purposes toward understanding the social and natural environment.

The curriculum content in the social and natural sciences should be organized around purposeful problems with a view to:

a) Making possible the use of a wide variety of stimulating educational material from literature, factual source books, visual and auditory

aids, with the local social and physical environment as a laboratory.

b) Making possible an appeal to many and varied interests.

c) Making possible the utilization of reading material with a wide range of difficulty, content, and interest appeal.

d) Stimulating and making possible a wide range of activities in reading, research, problem-solving, discussion, use of reference materials, writing and giving reports, letter-writing, organizing materials, planning, observing relationships, drawing conclusions, formulating generalizations, dramatization, understanding and using all forms of art, construction activities, using mathematics in a functional way, taking responsibility and cooperating in group projects and activities, all to the purpose of developing skills, understandings, ideals, values, beliefs, and attitudes.

e) Giving purpose to, and affording a use for, the basic skills in the language arts, reading, and mathematics. Skills in these areas must be given constant, definite, and systematic attention, but much practice in their use should come in the social and natural sciences.

2. The grade levels at which certain knowledge, skills, and abilities should be learned cannot be determined with any degree of specificity. It is common practice, for example, in the mechanics of capitalization, to specify that children should learn in the third grade to capitalize "Mr.," "Mrs.," and "Miss"; in the fourth grade, the names of cities, states, streets, and organizations; in the fifth grade, the names of persons and firms and the first word in a line of poetry; in the sixth grade, the names of the Deity, the Bible, and abbreviations of titles and proper names. Experienced teachers know that some pupils will learn all these by the close of the third grade; others will not have learned them by the time they graduate from high school. The teacher must be prepared to lead each child through the next steps in his development, regardless of the level he has achieved.

Similar graded lists of things-to-be-learned are provided in traditional instructional materials in all subjects: lists of words in spelling, lists of processes in arithmetic, lists of exercises in handwriting, lists of capitalization, punctuation, and usage rules in language, lists of skills to be developed in reading, and so forth. Properly used, these lists of essential learnings have value. Two values will be given attention. First, they may be considered as learnings to be introduced and emphasized when their purpose is clear. (The purpose of the most essential basic skills is clear even for the first-grade child.) They should never be considered as things to be learned in a one, two, three fashion, once and for all time, out of their functional setting and natural content, or as centers around which all in-

struction should be organized. The practice of organizing the curriculum almost entirely around these piecemeal, itemized goals has been the greatest limitation of the traditional school. If we learned to play golf this way, we should spend all our time with the instructor, drilling on itemized elements of the game, but never playing a game of golf. The purposes of drill in golf emerge from the game.

The second use to be made of such lists is in the systematic and economical diagnosis of individual deficiencies. At regular intervals the pupils should be tested for knowledge of essential spellings, essential processes in arithmetic, essential handwriting elements, essential English skills, and essential reading skills, the purpose being to keep both the pupil and the teacher constantly aware of individual deficiencies and developments. Such tests can be of the informal, teacher-made type, covering the spelling list for the grade at the rate of twenty words per week, mixed fundamental problems in arithmetic covering every process learned up to the time of the test, dictation exercises in English covering various aspects of the mechanics of writing, or they may be the more formal standardized materials such as tests of the various reading skills and diagnostic charts for handwriting.

3. Since life outside of school recognizes and rewards a great variety of aptitudes and combinations of aptitudes, the school should do the same. The traditional school has too often recognized and rewarded only docility and a facile memory. Teachers in such schools have been surprised when pupils they had considered hopeless achieved success in later life. The broadening of the elementary, high school, and college curriculums to include various forms of practical arts, fine arts, a school paper, athletics, extended educational field trips, participation in community affairs, stimulation of hobbies, participation in school government, the safety patrol, radio programs, work experience, and community health programs is evidence of the acceptance of this principle. The common school should be a proving ground in which the individual discovers his peculiar strengths and weaknesses. If every child is to find himself, the school's opportunities for development must be as broad as the demands of the culture, and the requirements must be within the limits set by the possibility of reality, purpose, and meaning for the individual child.

Functions of Measurement in the Diagnosis and Treatment of Learning Difficulties

Effective learning results in complex behavior patterns which may be differentiated into more or less related hierarchies of habits, skills, understandings, feelings, and desires. The product is infinitely complex, the

process of building it linear, and to a considerable degree sequential. The role of the school is to specify within the limits of individual capacities the behavior of educated people and then to determine the most effective sequences of experiences to bring it about. If sequence were not important, there would be little need for schools, since living in the complex culture would develop the necessary complex learnings. To make school experiences too lifelike would destroy the sequential organization of experiences essential to efficient learning. The relative importance of sequence and of the various criteria for determining it in the different areas of learning are points of issue between schools of educational thought.

In general, the criteria for determining optimum sequence of experiences are of two kinds: (1) those related to the physical, mental, and emotional maturation of the child, and (2) those related to the nature and complexity of the behavior to be learned. These are really different aspects of the same developmental process. Properly conceived, they both result in sequences which are challenging, purposeful, and meaningful to the learner. The "child centered" approach emphasizes the maturation process with a tendency to ignore goals, while the "culture centered" approach emphasizes the selection, refining, and grading of subject matter in the direction of definite goals. The first tends in practice to receive the greatest emphasis during the period of rapid maturation (primary and elementary), the second, as maturity is approached (high school and college).

That the development of motor, social, and intellectual abilities is highly sequential in nature is attested by the highly reliable age scales which have been constructed. The facts previously presented regarding the nature and extent of individual and trait differences are based on such scales. In general they indicate that individuals differ greatly in the rate of development of a given trait and in the level of development attained at maturity, also that the various traits of an individual develop at different rates and reach different levels at maturity.

Because of the waste and discouragement involved in attempting to teach again what the learner already knows or attempting to teach him at a level far beyond his present attainment, it is important that procedures be instituted for informing both the teacher and the learner of his status in a given sequence and what the nature of the next educational experiences should be if optimum development is to be achieved. This is true whether the curriculum is organized around itemized, piecemeal, socially validated goals in which sequence has been experimentally established, or around experience units in which functional relationships, meaning, and pupil-initiated activities determine sequence to a large degree. The variability of development in instructional groups is equally large under either system of cur-

riculum organization. Actually, both approaches have a definite place in curriculum development as long as the goals are seen in relation to the potentialities of the individual learner, to his probable future vocational status and needs, and when the teacher has a point of view and procedure which welcome and reward the IQ of 75 as well as the IQ of 125.

Whether the process of determining pupil status in a given developmental area and adjusting instruction to status should be called remedial teaching is questionable. It would seem to be just good teaching. It has been called remedial teaching because of erroneous ideas of the nature and extent of individual and trait differences and faulty conceptions of the schooling process based on them. The common criteria of need for remedial instruction have been: (1) discrepancy between measured intelligence and achievement in a given area, and (2) achievement status below grade status in a given area. The first criterion is based on the assumptions that individuals have equal aptitude in all areas and that this aptitude is measured by an intelligence test; the second assumes that all children should achieve up to the norms established for their age or grade group. Both of these assumptions we have found to be false. The best measure of what an individual should achieve in a given area is past achievement in that area. The need for remedial attention is indicated when progress in an area is stopped or markedly slowed down over a period of time.

The determination of pupil status in a given area of achievement and the adjustment of instruction to status should be a continuing process in every classroom. The role of testing in the process is an important one. Both achievement tests and diagnostic tests have important functions to perform. A general achievement test is one designed to express in terms of a single score a pupil's achievement in a given field of achievement. A diagnostic test is one intended to discover *specific* deficiencies in learning or teaching. It is a test in which a single total or composite score is of little or no significance, but on which the part scores or the percentages of correct responses to individual items are the measures sought. Tests may be diagnostic in various degrees. A test in English correctness, for example, may break the whole field up into such divisions as spelling, capitalization, punctuation, grammar, and usage, yielding a part score for each division, or may still further analyze each of these divisions, splitting the section on punctuation into tests on the use of the comma, period, semicolon, and so forth. Or it may make an even more detailed analysis, considering separately, for example, each of the types of situations in which, for instance, the comma may be used. To the degree, then, that the emphasis in the test is placed upon the part scores or upon percentages of responses to individual items, that test is of the diagnostic type. To the degree that the emphasis

is placed upon a single total score, designed to yield a measure of general achievement, the test is of the general achievement type. Perhaps the majority of the tests constructed for informal use by the classroom teacher are or should be of the *diagnostic type.*

The more general achievement test batteries which yield scores in vocabulary, reading comprehension, arithmetic reasoning, arithmetic computation, etc., have limited value in the planning of instruction for specific pupils. The major functions of such comprehensive batteries may be summarized briefly as follows:

1. To direct curriculum emphasis by:
   a) Focusing attention on as many of the important ultimate objectives of education as possible.
   b) Clarification of educational objectives to teachers and pupils.
   c) Determining elements of strength and weakness in the instructional program of the school.
   d) Discovering inadequacies in curriculum content and organization.
2. To provide for education guidance of pupils by:
   a) Providing a basis for predicting individual pupil achievement in each learning area.
   b) Serving as a basis for the preliminary grouping of pupils in each learning area.
   c) Discovering special aptitudes and disabilities.
   d) Determining the difficulty of material a pupil can read with profit.
   e) Determining the level of problem-solving ability in various areas.
3. To stimulate the learning activities of pupils by:
   a) Enabling pupils to think of their achievements in objective terms.
   b) Giving pupils satisfaction for the progress they make, rather than for the relative level of achievement they attain.
   c) Enabling pupils to compete with their past performance record.
   d) Measuring achievement objectively in terms of accepted educational standards, rather than by the subjective appraisal of teachers.
4. To direct and motivate administrative and supervisory efforts by:
   a) Enabling teachers to discover the areas in which they need supervisory aid.
   b) Affording the administrative and supervisory staff an over-all measure of the effectiveness of the school organization and of the prevailing administrative and supervisory policies.

However, such achievement test batteries are too general to be used as

a basis for instruction even when detailed analysis of items is made; although such an analysis has value and is certainly not to be discouraged. The sampling of items is too limited and the organization too gross for such tests to be considered as adequate guides in the planning and directing of educational experiences for individual pupils.

Two approaches have been made to the problem of determining status in a learning area as a basis for further instruction. One is commonly called "readiness testing" and is represented by reading readiness tests and arithmetic readiness tests at the elementary school level and by aptitude or prognostic tests in algebra, geometry, and language at the secondary school and college levels. Such tests are based on an analysis of the learnings essential to satisfactory progress in a predetermined instructional sequence. Too frequently the emphasis is placed on discovering pupils who cannot profit from a given course of instruction rather than on determining the optimum course for each pupil. That is, readiness is considered as something to wait for, rather than something that should be developed. Instead of adjusting the curriculum to the individual, individuals are sorted in terms of an inflexible curriculum.

The other approach to determining status in a learning sequence is labeled "diagnostic testing." Such tests are administered after a period of instruction to determine points of faulty or inadequate learning in a detailed and analytical manner with a view to correction. Superior teachers constantly carry on the process of checking learning through direct observation of behavior and informal testing. The values of expertly prepared tests for this purpose are that: (1) They are more thoroughly analytical than most teachers are able to prepare. (2) They make the teachers aware of the important elements, necessary sequences, and difficulties of the process. (3) They save the teacher's time and energy in diagnosis and leave more for individual remedial work. (4) They help the pupil recognize his learning needs by systematically emphasizing his errors. (5) Remedial procedures are usually suggested or provided which save the teachers time and also aid in systematizing the process.

In order that diagnostic tests may be most effective, they should meet the following general criteria: (1) They must be an integral part of the curriculum, emphasizing and clarifying the important objectives. (2) The test items should require responses to be made to situations approximating as closely as possible the functional. (3) The tests must be analytical and based on experimental evidence of learning difficulties and misunderstandings. (4) The tests should reveal the mental processes of the learner sufficiently to detect points of error. (5) The tests should suggest or pro-

vide specific remedial procedures for each error detected. (6) The tests should be designed to cover a long sequence of learning systematically. (7) The tests should be designed to check forgetting by constant review of difficult elements, as well as to detect faulty learning. (8) Pupil progress should be revealed in objective terms. Diagnostic tests frequently deal almost exclusively with the more mechanical aspects of a learning sequence and neglect the higher abilities requiring relational thinking and problem-solving. This is defensible in that efficiency in learning the more mechanical aspects leaves more time for the development of the higher mental processes.

The considerable proportion of elementary school time devoted to diagnostic testing and remedial teaching in the modern school is not generally recognized. Almost all the time devoted to the formal teaching of spelling and handwriting is consumed in the search for, and correction of, specific errors. In arithmetic, reading, and the mechanics of English many modern schools devote no less than one-fifth of the time allotted for the development of skills to finding the specific deficiencies of individual pupils and correction procedures. Many such tests are pupil-scored, and self-diagnosis is stimulated.

Perhaps the term "diagnostic testing" is more justified in its use in connection with the treatment of the more severe cases of maladjustment and arrested development which require insight and treatment of a more complex nature. Here the learning difficulties are more general, more complex, and more difficult to locate. The object of concern is not so much with what to learn as with factors limiting learning in general or in a specific area. Understanding is sought by studying the physical, intellectual, emotional, educational, and environmental factors to discover reasons for the maladjustment.

Functions of Measurement in the Motivation of Learning

A motivating condition has three functions in the learning process: (1) the *energizing* function, to increase the general level of activity and effort; (2) the *directive* function, to direct the variable and persistent activity of the organism into desirable channels; and (3) the *selective* function, to determine the responses which will be fixated and the responses that will be eliminated. Testing procedure properly conceived and executed places the control of the learning process within the educator's power as no other teaching device does. The three functions of a motivating condition are inherent in the test situation and are important criteria in the evaluation of measurement procedure.

Energizing function.—The extent to which examinations increase the general level of learning activity and effort is attested by the cramming sessions in high school and college which precede examination periods. In many schools the examination is the payoff. Success or failure depends upon them. The examinations determine to a large extent when students study, what they study, and how they study. Each instructor is scrutinized throughout his course to determine what he will emphasize and the type of questions he will ask in the final. Unless the examinations truly measure the real objectives of the course, the value of such motivation may be questioned. The use of such tests as a basis for relative grading in the schools attended by all children may be criticized in the light of what has been said previously in this chapter.

It has been reasoned that if tests increase effort, the more frequent the testing the greater the total effort. It is probable that the optimum frequency of testing depends upon the nature of the tests, the method of teaching, whether pupils score their own or each other's tests, the subject matter, and the ability of the students. When given too often, the relative importance of each test is reduced. Daily examinations probably have but little more energizing effect than daily assignments. It has been found (*18, 32, 37, 17*) in a variety of fields at the college level that when tests are given weekly and the results discussed, individual errors noted, and the final examination made up of similar questions, the lower-ability students achieve more than with less-frequent examinations. However, the more able students may be retarded by this process unless there are in each test, items of sufficient difficulty to challenge their ability. If, as is frequently true, the examinations cover only minimum essentials, it is possible they would profit more from the additional material which could be covered if less time were devoted to testing and remedial work. It is probable that the advantages of frequent testing even for the less able are the result of directing their learning and selecting the right responses rather than of an increase in energy stimulated by the examinations.

The extent to which knowledge of success on examinations motivates learning has been carefully investigated. In one of the earliest and most carefully controlled experiments Panlasiqui (*29*) administered twenty weekly fifteen-minute tests in mixed fundamentals of arithmetic to 358 matched pairs of fourth-grade pupils. In the experimental classes the amount of progress from week to week was stressed. Progress charts for both the individual and the class were kept, studied, and discussed. In the control classes no emphasis was placed on how much pupils scored or progressed. The difference between the achievement of the two groups

over the twenty-week period was statistically significant. Knowledge of progress was beneficial in relation to the amount of progress. The more able pupils profited most. Those who made little progress, of course, were little stimulated by it.

The influence of general praise and general reproof following an examination has been investigated by Hurlock (*16*). The results indicate that, in general, praise is more stimulating than reproof. However, praise is most effective with the immature and less able student, reproof is generally most effective with the mature and competent student. Either praise or reproof is more stimulating than no comment.

In general, the energizing effect of examinations depends largely upon the degree to which students are successful in them. Those who make high marks, who progress and receive praise are stimulated. Those who make low marks, who do not progress and are reproved become discouraged. The necessity of establishing reasonable educational goals, with at least partial success within the reach of all, seems indicated.

Directive function.—It is difficult to overemphasize the importance of examination procedure in determining what teachers teach and how they teach—what pupils learn and how they learn. The facts are well established.

In regional, state-wide, and local school testing programs where schools, teachers, and pupils are rated to some extent by the test results, the nature of the tests largely determines the quality of the schooling process. When tests are imposed upon the teacher and pupils, with important quality judgments involved, they become powerful instruments for determining educational goals and methods. The effects may be good or bad, depending upon the nature of the tests.

If the tests are based on traditional curriculum materials of the factual type, designed to determine the amount of textbook material which has been memorized, they have the effects ascribed to them by Douglass (*11*) of: (1) artificially determining objectives and method and of "freezing" the curriculum, (2) encouraging memorizing, cramming, regimentation, and mechanization, (3) reducing the teacher to the status of a tutor, (4) emphasizing only those outcomes that can be measured by objective tests, and (5) standardizing the curriculum, preventing adaptation to local needs and the further evolution of educational procedure. However, if the tests measure important study skills and problem-solving abilities, emphasize generalizations and their application to new situations, clarify important and neglected objectives, and focus attention on the ultimate objectives of education—the permanent learnings—then they may have the general effect of "thawing out" the curriculum of the schools, and stimulating more

acceptable teaching methods, objectives, materials, and learning procedures.

The influence of examinations on study procedures and objectives has been well established (*35, 12, 36, 26*). In preparing for essay-type examinations, students tend to outline and organize material systematically in large units, emphasizing relationships, trends, and personal reactions. In preparing for the broad sampling type of factual objective examinations, students emphasize factual details, names, dates, and results of specific experiments. A difference was found in the way students prepare for different types of objective tests. When preparing for completion tests, for example, students attempt a word-for-word mastery of important statements. In preparing for true-false tests, definitions and detailed facts tend to be emphasized. Differences in what is learned, the amount retained over a period of time, and the ability to take different types of tests have been experimentally related to the type of test a student prepares himself to take.

Certainly one of the most important criteria of the value of an achievement test is the degree to which it directs teaching and learning procedures into desirable channels resulting in the achievement of the most acceptable objectives.

Since comprehensive and systematic testing programs direct teaching procedures and learning toward the objectives measured, there is danger of warping the curriculum in the direction of measured objectives. The necessity of measuring all important objectives is indicated.

Selective function.—The extent to which tests help fixate correct and desirable behavior and eliminate errors depends not only upon the nature of the instrument but also upon how it is scored, and upon the emphasis placed on individual errors and remedial work in the follow-up procedure.

The selective function of measurement is related to the diagnostic function dealt with in a previous section of this chapter. In diagnostic testing emphasis is placed on diagnosis by the teacher or the educational psychologist. Here the concern is more with the student and teacher obtaining direct knowledge of errors (and, if possible, the reason for these errors) from the test situation.

Maximum learning results from testing when students are permitted to score their own papers and discussion of errors and remedial work follow immediately. Little (*23*) reports a well-controlled investigation involving the selective function of tests utilizing fourteen sections of a course in educational psychology averaging thirty students each. Four experimental sections took twelve unit tests during the quarter. The tests were immediately scored by machine, returned to the student, and the errors dis-

cussed in class. Four other experimental sections used a drill-machine placed on the desk of each student. The same twelve tests were taken on the machine. Each student indicated his answer to each question by pressing one of five keys. If the answer was correct the drum of the machine turned up the next question; if incorrect the same question remained before the student until the correct key was punched. The score was the total number of tries made in completing the test. The six control sections took only the pretest, mid-term, and final, which were administered to all sections. Both experimental groups showed superiority over the control sections on the final examination, with differences significant at the 1 percent level. The superiority of the drill-machine group was approximately twice that of the machine-scored group.

Logic and experimental evidence indicate that in the test situation the more immediate and direct the student's knowledge of when and why he is correct and when and why he is incorrect, the greater the tendency to fixate the correct responses.

Functions of Measurement in the Development and Maintenance of Skills and Abilities

When test items approximate the situations in life in which the learning will function, each item is not only an excellent test item, but a good teaching question as well. Such an examination becomes an effective learning device because it requires thought and problem-solving effort, and not mere recall or recognition of previous learning. Because of the additional motivation resulting from the test situation, it is probable that more learning takes place during the examination period than during any other equal period of learning time.

Recognizing the potentialities of tests for stimulating learning, especially in the skill and problem-solving areas (notably reading, mathematics, science, and the mechanics of English) series of drill material tests have been constructed as aids in developing and maintaining skills. Although the measurement function of these drill-tests is important from the standpoint of measuring progress and motivation, this function is really subsidiary to that of developing and maintaining skills and problem-solving ability.

The general characteristics of such drill-tests are: (1) the maintenance program is integrated with the development program; (2) the temporal distribution of drill is controlled to give a maximum of maintenance with a minimum of time; (3) each drill is graduated in difficulty; (4) progress records and/or charts are provided; (5) drill is provided on all aspects

of the skill; (6) specific difficulties reoccur at regular intervals; (7) diagnosis of error and remedial work are provided for; (8) several skills or abilities may be required by each test, but they are limited in scope and specific in nature.

Such materials have been available in the better-planned workbooks and in the form of self-testing drills for over twenty years. When teachers have insight into the nature of their construction and of their proper function in the learning process, they have been accepted as valuable teaching aids. To be valuable, the design of such material must be based on experimental evidence and adjusted to the ability of the student (several levels of difficulty should be used at a given grade level), and progress should be recognized in units more sensitive than age or grade scores. The danger is in overusing such materials to the extent that only the mechanics of a process are emphasized. When time is saved by these drill-tests in the more mechanical aspects of a skill area, it should mean that more time can be spent on reasoning, insights, understanding, applications, and the higher relationships.

Educational Measurement and the Improvement of Educational Practice

The variability in aptitudes of those we teach and the educational procedures suggested to secure a maximum of learning and development have been the burden of this chapter. To those who know well the schools, the teachers, and school budgets, the proposals may seem visionary and impractical. To those who know schools largely in the abstract, the recommendations may seem commonplace.

Much could be said and more should be known concerning variability in the competencies of teachers. One million among the sixty million of employed people in the United States are teachers. That this group is highly selected with reference to the qualities required for teaching is improbable. There is evidence that those enrolled in many teachers colleges represent the "mine run" of high school graduates in academic competence. Some colleges with students below this level have been found. It is probable that in the majority of classrooms above the primary level high-ability pupils will be found who exceed their teachers in general intelligence, reading comprehension, and problem-solving ability. In teaching, as in the other professions, there will never be enough high-level talent available to fill the positions.

The low relationship commonly found between the competence of a teacher in a given field of learning and the achievement of his students

in that field has served to salve our concern. But a more adequate analysis of such data indicates that, in general, the high-level students achieve the most with a high-level teacher and that low-level students achieve the most with a low-level teacher (*33*). This suggests that through measurement we may ultimately be able to place the student with the right teacher as well as give him the right book and set of problems.

The influence of the social climate of the classroom on academic achievement is still an area for investigation, but its influence on social behavior both inside and outside the classroom has been documented (*21*). It has also been shown that the social climate which a teacher habitually maintains in his classroom is highly related to the teacher's attitudes toward students, the school, fellow-teachers, and teaching procedures (*6, 20*). These attitudes have been successfully measured. By knowing a teacher's score on an attitudes inventory, it is possible to predict the social climate he will maintain in the classroom. The resistance of these attitudes to education and teaching experience indicates sufficient stability for purposes of teacher selection.

Although progress in the more appropriate selection, education, and placement of teachers may be anticipated, it will still be true that a high proportion of teachers will lack adequate insight into the educative process. The hope here must lie in improved materials of instruction. The administration and scoring of educational tests have been simplified to the level of a clerical task, but the interpretation and appropriate procedures that should follow testing require high-level ability and frequently extraordinary effort and determination. In most schools the needed instructional materials are not available. The teacher with the know-how is frequently forced to construct materials or to improvise with what can be found. For example, it is relatively easy to determine the reading comprehension levels of pupils in a sixth-year class, but to find materials of appropriate difficulty on the proper subjects with the proper interest appeal is extremely difficult and in most schools impossible. The schools have a long way to go in providing materials and equipment necessary for meeting the needs revealed by educational tests.

In summary, it may be said that educational measurement has taken a leading role in analyzing and refining educational objectives with reference both to their ultimate importance and to the most efficient and appropriate learning experiences. It has furnished the means of clarifying these objectives to both the teacher and the learner. It affords the means of determining the status of the learner in each learning area, suggesting appropriate experiences. It has emphasized the importance of the *how* in learning as

well as the *what,* placing study habits and intellectual skills in proper perspective. It has focused attention on child development and intellectual growth, diverting it from the subject-matter-to-be-covered point of view. It not only motivates the learner but also directs his efforts into effective channels. It furnishes the school official with a more adequate conception of his responsibility and serves as a guide to a more effective school organization and the selection of educational materials and facilities. Truly, educational measurement has outrun educational practice, but its leadership is wholesome and effective.

Selected References

1. ANASTASI, A. "Practice and Variability: A Study in Psychological Method," *Psychological Monographs,* **45**: 1–55, 1934.
2. ANASTASI, A., and FOLEY, JOHN P., JR. *Differential Psychology.* New York: Macmillan Co., 1949. Chaps. 14, 15.
3. BUCKINGHAM, B. R. "The Greatest Waste in Education," *School and Society,* **24**: 653–58, 1926.
4. BURR, MARVIN Y. *A Study of Homogeneous Grouping.* ("Contributions to Education," No. 457.) New York: Teachers College, Columbia University, 1931.
5. COOK, WALTER W. *Grouping and Promotion in the Elementary School.* ("Series on Individualization of Instruction," No. 2.) Minneapolis: University of Minnesota Press, 1941.
6. ———. "Measuring the Teaching Personality," *Educational and Psychological Measurement,* **7**: 399–410, Autumn 1947.
7. ———. "Some Effects of the Maintenance of High Standards of Promotion," *Elementary School Journal,* **41**: 430–37, Feb. 1941.
8. CORNELL, ETHEL L. *The Variability of Children of Different Ages and Its Relation to School Classification and Grouping.* ("University of the State of New York Bulletin," No. 1101; "Educational Research Studies," 1937, No. 1.) 98 pp.
9. COXE, W. W. *Levels and Ranges of Ability in New York State High Schools.* ("University of the State of New York Bulletin," No. 1001; "Educational Research Studies," 1932.)
10. ———. *Study of Pupil Classification in the Villages of New York State.* ("University of the State of New York Bulletin," No. 841; "Educational Research Studies," 1925.)
11. DOUGLASS, H. R. "The Effects of State and National Testing on the Secondary School," *School Review,* **42**: 497–509, 1934.
12. DOUGLASS, H. R., and TALLMADGE, MARGARET. "How University Students Prepare for New Types of Examinations," *School and Society,* **39**: 318–20, March 10, 1934.
13. HAWKES, H. E.; LINDQUIST, E. F.; and MANN, C. R. *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Co., 1936. Pp. 23–26.
14. HOLLINGSHEAD, A. D. *An Evaluation of the Use of Certain Educational and Mental Measurements for the Purpose of Classification.* ("Contributions to Education," No. 302.) New York: Teachers College, Columbia University, 1928.
15. HULL, CLARK L. "Variability in Amount of Different Traits Possessed by the Individual," *Journal of Educational Psychology,* **18**: 97–104, 1927.
16. HURLOCK, ELIZABETH B. "An Evaluation of Certain Incentives Used in School Work," *Journal of Educational Psychology,* **16**: 145–59, March 1925.
17. KEYS, NOEL. "The Influence on Learning and Retention of Weekly Tests as Opposed to Monthly Tests," *Journal of Educational Psychology,* **25**: 427–36, Sept. 1934.
18. KIRKPATRICK, JAMES EARL. "The Motivating Effect of a Specific Type of Testing Program," *University of Iowa Studies in Education,* **9**: 41–68, June 15, 1934.

19. LEARNED, WILLIAM S., and WOOD, BEN D. *The Student and His Knowledge.* ("The Carnegie Foundation for the Advancement of Teaching Bulletin," No. 29.) New York: The Foundation, 1938.
20. LEEDS, CARROLL H., and COOK, WALTER W. "The Construction and Differential Value of a Scale for Determining Teacher-Pupil Attitudes," *Journal of Experimental Education,* **16**: 149–59, Dec. 1947.
21. LEWIN, K.; LIPPITT, R.; and WHITE, R. K. "Patterns and Aggressive Behavior in Experimentally Created 'Social Climates,'" *Journal of Social Psychology,* **10**: 271-99, 1939.
22. LINDQUIST, E. F., *et. al. Manual for Administration and Interpretation of 1938 Iowa Every-Pupil Tests of Basic Skills.* Iowa City: Bureau of Educational Research and Service, State University of Iowa, 1938. P. 23.
23. LITTLE, J. KENNETH. "Result of Use of Machines for Testing and for Drill, Upon Learning in Educational Psychology," *Journal of Experimental Education,* **3**: 45–49, 1934.
24. MCCALL, WILLIAM A. *Measurement.* New York: Macmillan Co., 1939. Chap. 11.
25. MCNEMAR, QUINN. *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin Co., 1942.
26. MEYER, GEORGE. "An Experimental Study of the Old and New Types of Examination," *Journal of Educational Psychology,* **25**: 641–61, Dec. 1934; and **26**: 30–40, Jan. 1935.
27. NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. *Thirty-fourth Yearbook: Educational Diagnosis.* Chicago: University of Chicago Press, 1935. 563 pp.
28. ———. *Thirty-fifth Yearbook, Part I: The Grouping of Pupils.* Chicago: University of Chicago Press, 1936. 319 pp.
29. PANLASIQUI, ISIDORO. "The Effect of Awareness of Success on Skill in Arithmetic." Unpublished Doctor's dissertation, State University of Iowa, 1928.
30. PETERSON, J., and BARLOW, M. C. "The Effects of Practice on Individual Differences," *Twenty-seventh Yearbook of the National Society for the Study of Education, Part II: Nature and Nurture: Their Influence upon Achievement.* Bloomington, Ill.: Public School Publishing Co., 1928. Pp. 211–30 (chap. 14).
31. REED, H. B. "The Influence of Training on Changes in Variability in Achievement," *Psychological Monographs,* **41**: 1–59, 1931.
32. ROSS, C. C., and HENRY, LYLE K. "The Relation between Frequency of Testing and Progress in Learning Psychology," *Journal of Educational Psychology,* **30**: 604–11, Nov. 1939.
33. STEELE, LYSLE HUGH. "A Study of the Relationship between Teacher Proficiency in the Basic Skills and Pupil Achievement in the Same Skills." Unpublished Master's thesis, University of Minnesota, Aug. 1947.
34. TERMAN, LEWIS M., and MERRILL, MAUD A. *Measuring Intelligence.* Boston: Houghton Mifflin Co., 1937.
35. TERRY, PAUL W. "How Students Review for Objective and Essay Tests," *Elementary School Journal,* **33**: 592–603, April 1933.
36. ———. "How Students Study for Three Types of Objective Tests," *Journal of Educational Research,* **27**: 333–43, Jan. 1934.
37. TURNEY, AUSTIN H. "The Effect of Frequent Short Objective Tests upon the Achievement of College Students in Educational Psychology," *School and Society,* **33**: 760–62, June 6, 1931.
38. TYLER, R. W. "Permanency of Learning," *Journal of Higher Education,* **4**: 203–4, 1933.
39. WERT, J. E. "Twin Examination Assumptions," *Journal of Higher Education,* **8**: 136–40, 1937.
40. WOOD, BEN D. "The Need for Comparable Measures in Individualizing Education," *Educational Record,* **20**: 14–31, Jan. 1939.

2. The Functions of Measurement in Improving Instruction

By Ralph W. Tyler
University of Chicago

Collaborators: W. W. Charters, *formerly of Ohio State University;* Henry S. Dyer, *Harvard University;* Ruth E. Eckert, *University of Minnesota;* J. W. Wrightstone, *New York City Schools*

Since the purpose of this chapter is to outline the ways in which educational measurement, that is, achievement testing, can serve to improve instruction, we shall consider first what steps are involved in an effective program of instruction and then indicate the contributions that achievement testing can make to each of these steps. In this connection it will be noted that educational measurement is conceived, not as a process quite apart from instruction, but rather as an integral part of it. After discussing the contributions that achievement testing can make in the planning and conduct of an instructional program, the chapter then treats the ways in which educational measurement can contribute to effective supervision of instruction and to administration.

What Is Instruction?

The relationship between achievement testing and instruction can be seen more clearly by noting the nature of the instructional process. Basically, instruction is the process by which desirable changes are made in the behavior of students, using "behavior" in the broad sense to include thinking, feeling, and acting. Instruction is not effective, therefore, unless some changes in the behavior of students have actually taken place. Thus, in a certain English course one purpose of instruction may be to develop increased skill in writing clearly and in well-organized fashion; another purpose may be to develop increased ability to read and to interpret novels. English instruction in such courses will not have been effective unless these kinds of behavior changes have actually taken place in the pupils—unless the students do become more skillful in writing and are better able to read and interpret novels. To take a second illustration, instruction in a

certain science course may be aimed at developing an understanding of certain important science principles, an ability to utilize these principles in explaining common scientific phenomena, and some skill in analyzing scientific questions to see the kinds of data needed to solve them. Instruction in this science course will not have been effective unless changes of these kinds actually take place in the students.

Unless instruction is to be merely a haphazard or intuitively guided process, it requires rational planning and execution in terms of the plans. Viewed in this way, instruction involves several steps. The first of these is to decide what ends to seek, that is, what objectives to aim at or, stated more precisely, what changes in students' behavior to try to bring about. The second step is to determine what content and learning experiences can be used that are likely to attain these ends, these changes in student behavior. The third step is to determine an effective organization of these learning experiences so that their cumulative effect will be such as to bring about the desired behavior changes in an efficient fashion. Finally, the fourth step is to appraise the effects of the learning experiences to find out in what ways they have been effective and in what respects they have not produced the results desired. Obviously, this fourth step is educational measurement, or achievement testing. It is an essential part of instruction because without appraisal of the results being attained, the instructor has no adequate way of checking the validity of his judgments regarding the values of particular learning experiences and the effectiveness of their organization in attaining the ends of education.

In appraising the effects of the learning experiences today we not only test but also evaluate. "Evaluation" designates a process of appraisal which involves the acceptance of specific values and the use of a variety of instruments of observation, including measurement, as the bases for value-judgments. From the point of view of its functions it involves the identification and formulation of a comprehensive range of major objectives of a curriculum, their definition in terms of pupil behavior, and the construction of valid, reliable, and practical instruments for observing the specific phases of pupil behavior such as knowledges, information, skills, attitudes, appreciations, personal-social adaptability, interests, and work habits. Any learning situation has multiple outcomes. While the child is acquiring information, knowledges, and skills, there are also taking place concomitant learnings in attitudes, appreciations, and interests. This view indicates a shift from a narrow conception of subject-matter outcomes to a broader conception of growth and development of individuals.

Measurement Helps in Selecting Objectives

Educational measurement is not only the fourth phase of the instructional process, but it also contributes to the first phase, that is, the selection of the objectives to be sought. This contribution can be seen by examining the step of selecting educational objectives. Since it is possible to bring about a great many changes in behavior through the process of instruction, since some of these changes would be quite undesirable (for example, teaching children to steal), and since the time which the school can devote to instruction is relatively small and does not permit the attainment of all the possible desirable objectives, it would seem evident to an intelligent observer that every school and every instructor needs to determine what objectives shall be aimed at and to select a small enough number so that they can be attained with some degree of success. Although this argument for a careful selection of objectives seems obvious, as a matter of fact, a great many schools and teachers carry on the process of instruction without having a clear conception of the ends to be reached. It is true in schools and colleges as in other social institutions that forms, materials, and procedures become traditional and are passed on from one generation to another without a clear realization of the ends they are expected to achieve. Hence, many of us find ourselves using particular books and carrying on particular kinds of instructional procedures, not so much because we have certain ends in mind, as because these materials and these procedures have been used for many years and, possibly, for many generations and are accepted as of value without any clear perception of their underlying purposes.

Hence, people may carry on instruction without having clearly formulated the objectives. Yet, if instruction is to be rationally planned and choices are to be made among various possible materials and procedures of instruction, there must be a clear understanding of the ends to be sought.

The selection and clarification of objectives can be facilitated by carrying on a program of educational measurement. It is not possible to construct a valid achievement test, or to use one properly, without clarifying the objectives which the test is supposed to measure. One cannot measure the outcomes of a course in English without knowing what particular changes in behavior are sought in the English course since the test is a device for determining whether or not these changes have actually occurred. If the English course seeks to develop skill in organizing written material, then it takes a different kind of test than does a course which is aiming at developing a knowledge of certain types of literary materials or certain skills in reading, or certain feeling responses to novels. The necessity of

having objectives clearly formulated, that is, stated in specific terms rather than in terms of vague generalities, stimulates the instructional staff to attack the problem of objectives and to carry the analyses to a degree of definiteness and clarity that would not be likely to occur if the testing problem were not uppermost in mind. Hence, educational measurement can help in selecting and clarifying educational objectives by stimulating the faculty to formulate their objectives and to express them clearly in terms of behavior.

Selecting and defining objectives in operational terms does not in itself guarantee a wise choice of educational ends. Although the selection of goals on the part of school, college, or individual teacher is a matter of choice in the light of cherished values rather than a process of objective recognition, there are types of data that can be obtained by the school, college, or instructor that will provide bases for wiser decisions than when the choice of goals is made without such information. These include: (1) data regarding the students themselves, their present abilities, knowledge, skills, interests, attitudes, and needs; (2) data regarding the demands society is making upon the graduates, opportunities and defects of contemporary society that have significance for education, and the like; (3) suggestions of specialists in various subject fields regarding the contributions they think their subjects can make to the education of students.

It should be clear that the ends to be aimed at in a particular school or a particular course should be ends not already attained by the student, but goals that can be built upon his previous background of skills, abilities, knowledge, attitudes, and interests, and objectives that to some degree help the student to deal with his own problems, to satisfy his interests, to meet his needs. It is also clear that the ends of education in a particular school or course should be to some degree those knowledges, skills, abilities, attitudes, and interests that have significance in contemporary society, and help to carry on an effective civilization and improve it. Hence, to provide suggestions about objectives like these, data regarding the students and about contemporary society are both helpful. Furthermore, specialists who have thought and worked intensively in a particular subject field sometimes get significant insights regarding the contribution their field can make to the education of children or youth. These suggestions are most likely to be found in reports of committees on the curriculum or teaching of the several subjects. These reports can be helpful in suggesting objectives.

The three kinds of data listed will suggest more objectives than any one school or course can possibly attain. Furthermore, some of these objectives will be conflicting in their nature. It is necessary to make a choice among them. In making such a choice, the philosophy of education held

by the school or the instructor serves as a guide to identify those objectives of greatest value in terms of the conception of the "good life" and the "good society" implied by this philosophy. The school will wish to emphasize objectives that are most in harmony with its philosophy.

Another consideration in choosing objectives is the findings of studies in the psychology of learning. Objectives that are not likely to be attainable according to the psychology of learning will be omitted as will objectives that are more appropriately developed at some other stage in the students' maturation.

The foregoing is a brief outline of the procedure involved in a rational process of selecting objectives. In this process, educational measurement serves particularly by providing data about the students that suggest educational objectives. For example, in formulating objectives for a ninth-grade course in mathematics, all the students completing the eighth grade were tested in terms of their arithmetic skills, their ability to analyze and to solve a variety of commonplace problems involving quantitative aspects, and their understanding of certain basic mathematical concepts. The weaknesses shown in problem-solving and in understanding mathematical concepts suggested that these were objectives to be emphasized in the ninth-grade course.

To cite another example, the instructors in a freshman college course in music gave a battery of tests during the freshman orientation week to get suggestions regarding objectives for the music course. The tests involved ability to recognize similar musical themes, to identify types of musical instruments in various parts of orchestral records, to distinguish changes in pitch and in rhythm; they also involved knowledge of musical history and familiarity with various classical and popular works. Relatively low scores made by students on tests assessing recognition and understanding of musical recordings suggested the need for emphasizing as an objective "the development of ability to hear and interpret music."

These two examples are cited here to illustrate the use of educational measurement as a part of the process of rational selection of educational objectives. Educational measurement thus contributes in two ways to this first step in the process of instruction. It stimulates the instructional staff to select and to define its objectives clearly and it provides a tool of value in obtaining data to be used in making a wise selection of objectives.

Measurement Helps in Selecting Content, Learning Experiences, and Procedures of Instruction

The second major step in the process of instruction is to select content and learning experiences and to plan procedures of instruction that are

likely to attain the ends sought, that is, the objectives that have now been selected and clearly defined. Educational measurement can contribute to this step in several ways. In the first place, the results of achievement tests given at the close of the preceding grade or course or at the beginning of the current one provide a basis for judging the various levels at which students are ready to proceed toward the attainment of each major objective and, therefore, the several levels of content and the kinds of instructional materials and procedures likely to be effective. For example, if one important objective of the English course is to develop ability to organize written materials effectively, a test of ability to organize given at the beginning of the course or at the close of the preceding course will indicate to what degree the various students have already attained some skill in organization and whether many of them are now ready to proceed with more complex organizational problems or should begin with much more elementary organizational problems in their writing. If another objective of the English course is to develop the ability to interpret literary material with skill and understanding, tests given at the beginning will indicate what kinds of interpretations the students are now able to make and, therefore, what kinds of materials are appropriate for them in developing further skills and understanding in interpretation. Similar illustrations could be drawn from other fields. For example, in the case of a certain science course, if the objective is to develop some ability to apply important principles of science to common concrete phenomena, an examination given early in the course should indicate what principles the students already understand and to what kinds of concrete problems they can apply these principles. It is possible from the data to make inferences regarding the range of the appropriate content to be used with this group of students and the various instructional procedures that are likely to be effective with them.

The second way in which the results of achievement tests can be used to guide the selection of content and procedures of instruction is through the evidence they furnish regarding the effectiveness of particular content and procedures that have been tried out experimentally. When these experiments are properly designed and controlled, the results provide some basis for judging the kind of content and the kind of instructional materials which are effective means for attaining the instructional goals. The results of the tests also serve as criteria for testing the hypotheses upon which the selection of content and instructional procedures was based. In some cases the results may indicate that the materials and procedures are highly effective. In other cases students may show little or no progress toward the educational goals. In most cases there will be some respects in which satisfactory results are being attained and other respects in which results

are far from satisfactory. In the latter cases, such results of testing provide a basis for restudy, replanning, and modification in the content and instructional procedures. Detailed analyses of the test results may help to indicate whether the failure to attain the results desired is due primarily to poor material or poor instructional procedures, or whether the very principles or hypotheses upon which the instructional plan was based may be invalid. This use of test results provides a rational basis for the continued revision and improvement of the content and learning experiences.

An example may illustrate this contribution of measurement more concretely. A certain course in science had as one of its aims the development by the student of the ability to use science principles in explaining some of the common natural phenomena. As the course was first conducted, an effort was made to include materials which described a variety of natural phenomena and to give the student a chance through field trips and laboratory work to come in direct contact with many of these phenomena. It was found by testing as the course went on that the students gained a great deal of information about concrete phenomena, but were not able to explain them in terms of the appropriate scientific principles. Since the results were disappointing in every unit of the course, it seemed likely that the poor results could not be attributed merely to a particularly poor selection of materials or a particularly poor field trip or experiment. Rather, the results raised a question as to the validity of the hypothesis upon which the course was planned, namely that familiarity with natural phenomena and descriptions of them would provide the means for students to learn to explain these phenomena in terms of scientific principles.

In case a basic hypothesis upon which instruction is planned is brought into serious question, it is desirable to compare results secured with those obtained when a different hypothesis is used. In this example, the instructors tried out a second hypothesis, namely, that formulating scientific principles inductively and practice in using these principles in explaining natural phenomena would provide the means for students to learn to explain other natural phenomena in terms of scientific principles. Using this second hypothesis, the instructors selected material which required certain observations and stimulated the students to develop scientific principles inductively. Opportunities were also provided for students to apply these principles to concrete phenomena and to criticize and revise the explanations which they developed. Further tests on the basis of the revised instructional program indicated that there was a considerable improvement in the degree to which students were able to utilize scientific principles in interpreting the concrete phenomena around them. This is one of many possible illustrations of the use of educational measurement in checking

on content and learning experiences, and, when necessary, revising them. Achievement testing at the beginning and at the end of a period of instruction thus contributes to the effective selection of instructional content and learning experiences.

Measurement Helps in Organizing Learning Experiences

In organizing learning experiences, the purpose is to put together the various learning experiences in such a way that the cumulative effect is made much greater than that which would result from a haphazard organization of them. We may note that the organization of learning experiences involves relationships both vertically, that is, from one week or one month to the next within the same subject or field, and also horizontally, that is, from one subject to another and from the school sector of the child's experience to the out-of-school sector within the same period of time. Many learning experiences are relatively ineffective as isolated aspects of a student's life, but when they are appropriately organized so that one experience builds upon another and reinforces the other, profound changes in the student's behavior result. Since instruction is generally concerned with producing significant changes rather than merely transitory modifications in behavior, it is very important to devise schemes of organization which will increase the efficiency of the learning experiences and result in the maximum cumulative effect.

From this brief definition of the problem of organization, three criteria emerge for an effective organization of instructional experiences, namely, continuity, sequence, and integration. Continuity refers to the provision of continuing emphasis upon the desired knowledge, skills, abilities, attitudes, and the like over the months and years. For example, an effective curriculum in arithmetic will provide continuous opportunity for arithmetic computations so that these skills will not be developed and then permitted to wither away through lack of use. Similarly, a good curriculum will provide continuing opportunity to use and apply such a basic concept as "transformation of energy" so that it will become part of the student's way of thinking about the physical world.

The criterion of sequence refers to that arrangement of learning experiences in which each subsequent treatment of a particular concept, skill, attitude, or the like is not simply a repetition of former treatments, but deals with it more broadly or more deeply or at a higher level than the preceding ones. For example, sequence in the treatment of the concept of "cooperation" might begin with the idea of cooperation among a small intimate group including reference to physical needs, such as cooperation among family members in preparing and serving a meal. Subsequent treat-

ments of cooperation could broaden the groups in which cooperation is involved, increase the types of cooperative relationship, and deepen the intellectual and emotional significance of cooperation. This would imply real sequence rather than mere continuing repetition of the original concept of cooperation.

Correspondingly, sequence in the treatment of a skill like "map-reading" might begin with a simple chart which the pupils drew of their own classroom. As they learned how to interpret this chart, subsequent experiences with map-reading might involve a wide range of kinds of maps and a broader use of maps to make many kinds of interpretations.

The criterion of integration refers to a relationship among the several subjects or parts of the curriculum as well as between school experiences and those out-of-school, all of which serve to provide mutual re-enforcement for the important learnings emphasized in each of the several segments of the pupil's learning experiences. For example, the criterion of integration is involved when the attempt is made to provide opportunities in social studies and science for students to use the writing skills they are developing in the English class. The use of these skills in other classes helps to strengthen and render more meaningful the learnings gained in the English class. Likewise, attitudes of enjoyment of music developed in the home can be intelligently re-enforced by school experiences with music. These re-enforcements through continuity, sequence, and integration are very necessary in order to achieve an efficient organization that will provide a maximum cumulative effect of the various learning experiences of each student.

To plan for an effective organization of learning experiences, it is necessary to identify the instructional elements to be organized and to develop some hypotheses or principles that provide a basis for arranging these elements to produce the desired cumulative effect. Thus, in the case of a mathematics course it is necessary to identify the elements that are to be organized in the mathematics course. These are likely to be of two kinds—the first, concepts, that is, basic ideas or notions about mathematical matters, and, second, skills, that is, the important intellectual operations that are involved in handling quantitative computations and problems. An illustration of a mathematical concept might be the idea of place in the number system; another might be the concept of continuity or discontinuity in a mathematical function. An illustration of a skill might be the skill of multiplying simple integers. Each of these is an element which can be introduced at a relatively simple level and carried on through increasingly complex and broader aspects. Hence, both concepts and skills are elements which permit of effective organization. Correspondingly, in a field such as

social studies, the organizing elements are likely to include: (1) Concepts, such as the interdependence of all human beings. This concept can be introduced at a relatively simple level to small children as they see the degree to which they are dependent upon each other for various activities in the kindergarten and can be extended to increasingly broader and deeper ranges until mature students see the degree to which each citizen and each nation is dependent upon citizens and nations far removed from them. (2) Skills, such as skill in reading and interpreting social statistics and the like. These skills, too, can be introduced at a simple level and carried on to more complex operations. (3) Values, such as belief in the dignity and worth of every human begin. Each of these elements permits of appropriate organization to provide for continuity and sequence and to relate it more integrally with the rest of the student's experience, within and without the school.

In addition to selecting the elements to be organized, it is necessary to make some decision about the principles of organization to be followed. Thus, in connection with the development of such a skill as addition, the principle of organization may be to begin with what is apparently simple and move on through what appears to be increasingly complex. This is accomplished by beginning with the addition of two one-digit numbers, then proceeding to the addition of several one-digit numbers, then to the addition of two-, three-, and four-digit numbers, then to fractions, and ultimately to negative and imaginary numbers. The acceptance of such an organizing principle suggests the way in which this skill of addition is to be developed over the years. There is also the problem of the organizing principle to relate addition to other aspects of experience, for example, relating addition to subtraction, to multiplication, to division, and ultimately relating addition in mathematics to operations or concepts in other fields. Before a rationally planned program of instruction can be developed, it is necessary to decide upon elements to be organized and possible principles of organization to be applied.

Educational measurement can contribute to the process of organizing instruction in two ways. In the first place, the development of achievement tests for a particular course or instructional program requires an explicit statement about the elements of organization and the hypotheses which are accepted regarding the principles of organization. These explicit statements are necessary because the achievement test must be built so as to include these elements as aspects of the test. For example, in the case of mathematics, if concepts and skills are the two major kinds of elements, it is necessary for the achievement test to include items involving concepts and items involving skills. It is also necessary to indicate the principles of

organization so that the test can have varying levels of items, some involving concepts at a relatively simple level and others at an increasingly broader and deeper level. Similarly, items involving skills will have some at a relatively simple level and others at the more complex level, the definition of simple and complex demanding some formulation of the organizing principles actually used in developing this mathematics course.

Correspondingly, in the case of a social studies course it is necessary to identify the types of elements to be sure that the achievement test properly samples each of the types. If the social studies course involves concepts, skills, and values, then it is necessary that test exercises test for the understanding of certain concepts, for competence in certain social studies skills, and for the kinds of values which the students cherish and accept for themselves. It is also necessary to know the organizing principles for each of these elements, so that the test may provide opportunity for the students to show their ability to deal with increasingly more mature concepts, skills, and values. It is true in the case of organization, as in the case of objectives, that although a rationally planned instructional program requires a clear understanding of organizing elements and of organizing principles, many teachers have not given explicit consideration to these matters because they have accepted the traditional organization without examining the basis upon which such schemes of organization rest. Hence, it is quite possible for teaching to go on without the teacher being conscious of the scheme of organization and the elements of organization in the program he is teaching. The fact that a comprehensive achievement test cannot be constructed without some specification of these organizing elements and organizing principles stimulates the instructional staff to formulate their conceptions of the bases of organization and thus help to bring problems of organization more clearly to the attention of instructors. This is a significant contribution, because it is not likely that organization will be made more effective unless it is clearly and explicitly recognized by those who are responsible for instruction.

A second way in which educational measurement can contribute to the organization of instruction is by providing a means for testing the hypotheses about organization around which any given instructional program has been developed. For example, a certain social studies program was organized in the elementary school on the hypothesis that each of the major concepts would be introduced by beginning with the life in the family, then progressively extending these understandings to the school, the community, the state, the nation, and to the world community. On this basis such concepts as interdependence, human variability, political power, and the like were developed over a six-year period. To appraise the effectiveness

of this organizing scheme, tests were constructed for these concepts at several levels in the sequence. It was found that a number of the concepts did not develop very satisfactorily when they were extended in this geographic sense from the family to wider and wider circles, ultimately reaching the international scene. This brought the organizing principle into question and led to its reformulation. The revised organizing principle was based on the idea that certain aspects of life are simpler and easier to comprehend than other aspects and, hence, that the concepts would be developed first in their relation to these simpler aspects of life and then move on to other more complex aspects. The second scheme of organization was also appraised by the use of tests and was found to give somewhat better results than those obtained by the earlier scheme.

As another illustration of the use of educational measurement in making an explicit test of hypotheses about organization, we may cite the case of the English course which was organized with the idea that integration could be achieved by having the English course deal with literary writing and asking the teachers of science, social studies, shop, and mathematics courses to take responsibility for applying the concepts and skills in writing to their own particular fields. This principle of horizontal organization was appraised by giving tests that showed the degree to which students were able to utilize major skills in writing in connection with each of the areas of instruction. It was found that this scheme of organization was not totally satisfactory. In some cases the concepts developed in the literary field were not directly applicable to the writing of scientific papers or of papers in other fields, and in other cases the application required more understanding of the principles of writing than the teachers of the other fields had acquired. These results led to the modification of the scheme of organization so as to provide the initial development of writing principles and skills in the English class in connection with each of the types of writing commonly employed in other courses students were taking. After the initial development in the English class, the other classes provided for continued practice of these writing skills through periodic papers and reports. Judging from the test results, this organizing scheme proved to be more effective than the one previously used. These illustrations are but two of many that could be cited to show the way in which educational measurement is used to test the organizing principles of the curriculum.

The foregoing discussion has indicated two ways in which achievement testing can contribute to the process of organizing learning experiences. In the first place, the development of a valid testing program challenges the staff to consider the problem of organization and requires an explicit

formulation of the organizing elements and the principles of organization. Educational measurement also provides a means for testing these organizational hypotheses so as to retain those that are effective and to discard or reformulate those which do not produce the cumulative effect which good organization demands.

Measurement Aids in the Supervision and Administration of Instruction

This chapter has thus far dealt with the way in which educational measurement contributes directly to the process of instruction. It has been noted that achievement testing is in itself one of the major phases of instruction and that in addition it helps the formulation of clear-cut educational objectives, it provides assistance in the selection of content and learning experiences, and it aids in the development of an effective organization of learning experiences. Next, the discussion deals with the contribution educational measurement can make to the supervision of instruction and to its administration.

Measurement Contributes to the Education of Teachers in Service

Essentially, the supervision of instruction has two major functions. One is to provide for the continuing development of the teacher—in-service education—and the other, to provide for the coordination of instructional efforts. In providing ways by which teachers can improve their own knowledge and skill, achievement testing has two important contributions. In the first place, to work upon the selection or construction of appropriate achievement tests is an important kind of learning activity for the instructional staff. It requires a careful formulation of objectives and clear thinking about the meaning of these objectives so that they can be defined explicitly in terms of the actual behavior changes desired in the students. It also requires careful consideration of the ways in which one's students might be expected to display the kind of behavior which instruction aims to develop, and this helps to clarify the teacher's instructional tasks and keep him from being subject-centered. It has often been maintained that teachers should be students of children rather than merely students of subject matter. The kind of focus of attention involved in the construction and selection of achievement tests is one means of helping teachers to study children. It requires continual thought about the potential reactions of children, the changes to be desired in their behavior, the ways in which these changes can be brought about, the ways in which this changed be-

havior might be expressed within and without the classroom, and the ways in which it could be appraised validly. Surely this is an important contribution in the continued education of the instructional staff.

Another important contribution to the education of the instructional staff is provided when the results of achievement tests are used by the supervisor and the instructional staff in evaluating the instructional program. When it becomes clear that a particular course brings about some of the desired changes in behavior but not others, it raises the question as to what is wrong with the course if it does not attain certain of the purposes desired. This focuses attention upon possible modification of content at those points where the course needs to be improved, and it stimulates proposals by teachers for experimentation. As experiments are initiated, achievement tests again provide the means for determining whether a given experimental program is more effective than the program previously followed. Such use of test results stimulates further development of understanding and skill on the part of teachers by giving them an effective means of investigating instruction.

The use of a measurement program as a major supervisory device has one marked advantage over the sole dependence upon supervisory observations of teachers at work and later discussion of the observations. It focuses the attention of the teacher and the supervisor upon the students and their achievement rather than upon the teacher and his procedures. The older methods of supervision, which largely depended upon classroom visitation and subsequent conferences during which the supervisor criticized the materials and procedures used by the teacher, frequently caused embarrassment, resentment, or defense reactions on the part of the teacher. He often felt that the supervisor's criticism represented only personal opinion and that the teacher's judgment about particular material or a procedure was as valid as the supervisor's. He also felt ill at ease because the supervisor was looking at his activities closely and critically. Under these conditions neither the instruction observed nor the conference that followed were as satisfactory as might be desired. On the other hand, if the focus of supervision is upon the study of test results, the question becomes, "How does Johnny (or how does the class) react to this kind of material?" When defects are identified, the next questions become, "Why is it that Johnny (or Frank or the class as a whole) has difficulty with this kind of material?" "What can be done about it?" Instead of putting the teacher on the defensive, such questions are likely to cause him to be as much interested as the supervisor in noting difficulties and figuring out ways of overcoming these difficulties. This kind of objective focus of attention in supervision is likely to be more conducive to teacher cooperation than is

supervision which directs its attention to the teacher rather than to the students involved. This is a valuable characteristic of the measurement program as a supervisory device.

Measurement Helps in the Coordination of Instructional Efforts

In addition to in-service education another primary function of supervision is coordination. Education in school or college involves a complex program using a number of different people, teachers of different subjects, teachers of different grade levels, supervisory and administrative officers, all presumably working toward certain common ends that are involved in inducting young people into responsible adulthood in our society and achieving desirable control of their current problems. Any kind of complex program that uses a number of different people all working toward a common end is likely to get out of adjustment, to be uncoordinated. The function of coordination is to see that these various persons and groups make an effective contribution to the common purposes of the program.

Coordination of a professional group is not as simple as it appears to be in coordinating a large number of workers on an assembly line. Many decisions, many adjustments in a professional program are made as the program develops, and cannot be predicted with precision in advance. Instructional supervision is a cooperative activity involving a good deal of group planning and sharing of ideas. This kind of coordination requires two-way communication rather than directives sent out by the supervisor. Essentially instructional coordination involves (1) sharing common purposes on the part of all staff members, (2) an increased understanding of how to attain these common purposes so that each staff member sees his own job in relation to these purposes, and (3) on the part of each staff member, an increased understanding of himself and how to use his particular competence in the program to attain the common purposes.

Educational measurement contributes to coordination in several ways. Because an achievement testing program must begin with agreement upon objectives and a clear formulation and definition of them, this process in itself is one way of getting common purposes shared by all staff members and having these common purposes clearly understood. Furthermore, as the content and learning experiences are appraised by use of tests, each teacher gets a clearer view of the kinds of materials and procedures used by other teachers and their effectiveness. As the organization of learning experiences is appraised by testing, each teacher gets a clearer view of the kind of contribution each field and grade is making to the common purposes of the school. The study of test results to see what changes take

place in students in relation to the different courses they are taking provides a way for the teacher to see his own job in relation to the work of others and both in relation to the objectives of the school more clearly than is provided by general statements of the expected division of labor. Such study also makes some contribution to the teacher's understanding of his own competence and how it can contribute to the program at hand. Test results may show that teacher A has brought about marked changes in objectives 1 and 3, whereas teacher B has been more effective in connection with objectives 2 and 4. As these results are studied and questions are raised about their meaning, it becomes possible for the supervisor to indicate more clearly the kind of contribution which each teacher is best prepared to make and to suggest ways in which his own particular competence can be used even more effectively in helping students develop in desired ways. The important thing about the measurement program is that it focuses attention upon objectives and their attainment, which is the fundamental basis for coordination, that is, the relation of means to the essential ends of education.

Measurement Helps in Administration

Educational measurement can make a contribution to administration as well as to instruction and supervision. Administration has eight major functions: (1) the provision of personnel to carry on the instructional program, (2) the organization of the staff and effective relations among members of the staff and between members of the staff and appropriate groups outside the school, (3) the development of policy and plans for the educational program, (4) the provision of supplies, equipment, buildings, and the like that are needed for the work, (5) the creation of conditions requisite for effective work, (6) the provision of funds to carry on the work, (7) the evaluation of the work as a whole and its various parts, and (8) the interpretation of the work and needs to the public and to special groups whose understanding is essential for the adequate support and development of the educational program.

In examining this list of functions, it appears most obvious that educational measurement contributes to the evaluation of the work of the school and its various parts. Measurement also contributes to several other administrative functions. In selecting personnel, scores on achievement tests represent one index of competence. Furthermore, the quality of work the instructor does is partly indicated by the measured progress made by his students. Hence, educational measurement provides two, among several, indices helpful in selecting and promoting the instructor. The results of achievement testing make not more than a minor contribution in appraising

the effectiveness of staff organization. However, in the development of policy and the planning of program the results of achievement tests help to indicate particular points of emphasis needed in a particular school, and the process of planning for achievement tests provides, as has been pointed out earlier, an important basis for getting common understanding of policy and common agreement about the ends of the educational program. The results of achievement tests also provide an important basis for identifying instructional problems needing attack and for determining the effectiveness of efforts at instructional improvements.

The results of achievement tests provide helpful evidence that can be used in the selection of effective instructional materials and facilities. Buildings have too often been built primarily as monuments to the architects. Classrooms have commonly been planned in terms of convenience of construction or the patterns and traditions of the building industry. Instructional facilities are often provided on the basis of advertising claims and the whims of purchasing departments. All of these materials are primarily means for bringing about the objectives of education and can be selected more wisely and more validly on the basis of evidence as to their effectiveness in promoting the objectives of education. Carefully planned programs of achievement testing can provide data regarding the probable contribution of various kinds of buildings, facilities, supplies, and equipment that are more dependable than the usual bases for making decisions on these matters.

The results of a carefully planned program of educational measurement are also useful in appraising the conditions set up to promote staff growth and development. One important criterion of staff development and staff efficiency lies in the educational results the staff is achieving. These results can be inferred from the results of a carefully planned program of testing of the students' progress.

Measurement makes no direct contribution to securing funds except as it contributes to public relations. The interpretation of the work and needs of the school to the public and special groups whose cooperation is essential can be much more validly made when evidence is available as to the results being obtained and the difficulties being encountered than when the content of public relations is based on enrollments alone or on matters which do not directly reflect the desired results of the educational program. There was a time when American education was taken on faith, when it was assumed by the public that mere enrollment of the child in school was a good thing. Increasingly, however, the lack of evidence regarding the school's effectiveness has made it possible for public criticism of education frequently to go unchallenged. If the school is to meet criticism intelli-

gently, if it is to inform the public regarding its real achievements and its real difficulties, it must base its reporting not only upon a description of the educational program and upon the statistics of the students enrolled, but it must also indicate the kinds of results being achieved. It must have the data from programs of educational measurement to show clearly what the school is now accomplishing, where it is having difficulty, and what kinds of further support it needs to do a better job.

It can thus be seen that educational measurement has a contribution to make to administration at several important points in addition to the obvious contribution that it makes in providing the basis for a careful appraisal of the work of the school. As administrators recognize the many values accruing from a carefully planned testing program, educational measurement can make its potential contribution more widely effective.

Measurement Can Help Instruction More When Certain Conditions Are Met

Many potential contributions that educational measurement can make to the improvement of the content, organization, supervision, and administration of instruction have been suggested in the previous sections of this chapter. At present these potential contributions are rarely achieved in full measure. In order to realize them, certain important conditions are necessary in organizing and developing an achievement testing program. Among these, three are of particular importance. First, it is essential that the outcomes selected for testing should be the important objectives of the instructional program. For example, if the instructional program emphasizes primarily the development of certain intellectual skills and the acquisition of some major concepts, whereas the testing program concentrates upon the memorization of isolated and specific information, the testing program is either disregarded as being relatively unimportant or it tends to deflect the educational program from its primary objectives. The testing program must, therefore, provide a measure of the degree to which the important objectives of the instructional program are being attained.

In the second place, it is essential that achievement testing be planned and developed as an integral part of the program of curriculum and instruction. The tendency in some schools to set up a test division separate from the division of instruction is unwise. The tests should be based upon the objectives of instruction; the results should aid in the improvement of instruction. Hence, only insofar as tests are selected or constructed in terms of the instructional program and the results are immediately available for instructional planning can their greatest values be obtained. If achievement testing is not an integral part of curriculum and instruction, it is likely

to be viewed with suspicion, or to go off in directions other than those planned in the instructional program, or to operate as a parallel activity but not one influencing instructional development.

A third condition for realizing the potential contributions of measurement to instruction is that the achievement testing program be guided and controlled by those responsible for the guidance and control of the instructional program. Because achievement tests provide such a concrete indication of objectives, because they have great influence upon students and teachers in directing their efforts, it is possible for the achievement tests to be more influential in the actual learning that goes on in the school than the curriculum outline or any other materials prepared by the instructional staff. Hence, if the testing program is not under the direction and control of those responsible for the instructional program, measurement will operate as a powerful influence in directing learning in the school which is not in harmony at all points with the instructional program of the school.

A fourth condition for realizing the potential contributions of measurement to instruction is to encourage teachers to take part in the construction of all types of tests, since many standard tests and teacher-made testing instruments constructed for one school system are not applicable without adaptations to the curriculum of another school system. It is necessary for interested persons and teacher committees to work upon the construction of evaluation techniques designed to meet the special needs of their own school curriculum and courses of study. These teacher committees may work informally with the school's division of tests and measurements. For example, a committee of five chemistry teachers might work upon a battery of tests designed to measure various aspects of thinking in the subject of chemistry. They could develop subtests such as obtaining facts about chemistry from various sources, interpretation of facts and data in chemistry, and the application of principles of chemistry to new situations. Perhaps a committee of mathematics teachers working cooperatively with the school's division of tests and measurements could develop tests for the objectives of mathematics. This committee is strategically placed so that they can devise a test which is adapted to the school situation of which they are a part. The committee of mathematics teachers might embark upon some of the more challenging test areas, for example, understanding quantitative relationships, critical thinking in mathematics, and interests and attitudes related to mathematics.

Many of the present inadequacies of instruction are due to the common tradition in many schools of making no significant changes in the instructional programs and procedures over the years, or the equally unintelligent practice in many other schools of making changes on the basis of current

fads or the personal whims of the staff. The responsibilities expected of educational institutions in our times are wholly impossible of attainment unless great improvements are made in the effectiveness of instruction. Educational measurement can have a profound influence in the improvement of instruction; but to do so, it must be viewed as an integral part of instruction, its planning must go hand in hand with instructional planning, and the results must be used continuously to guide the planning and development of the curriculum.

Selected References

1. ADMINISTRATIVE OFFICERS OF PUBLIC AND PRIVATE SCHOOLS, Proceedings of the Ninth Annual Conference of. *Evaluating the Work of the School.* Edited by William C. Reavis. Chicago: University of Chicago Press, 1940.
2. BROWN, CLARA M. *Evaluation and Investigation in Home Economics.* New York: Appleton-Century-Crofts, 1941.
3. CLARKE, H. HARRISON. *The Application of Measurement to Health and Physical Education.* New York: Prentice-Hall, 1945.
4. *Cooperation in General Education: A Final Report of the Executive Committee of the Cooperative Study in General Education.* Washington: American Council on Education, 1947.
5. DOUGLASS, H. R. (ed.). *The High School Curriculum.* New York: Ronald Press, 1947.
6. DUNKEL, HAROLD B. *General Education in the Humanities.* Washington: American Council on Education, 1947.
7. GREENE, HARRY A.; JORGENSEN, ALBERT N.; and GERBERICH, J. RAYMOND. *Measurement and Evaluation in the Elementary School.* New York: Longmans, Green, 1942.
8. ———. *Measurement and Evaluation in the Secondary School.* New York: Longmans, Green, 1943.
9. HAWKES, HERBERT E.; LINDQUIST, E. F.; and MANN, C. R. *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin, 1936.
10. HORTON, CLARK W. *Achievement Tests in Relation to Teaching Objectives in General College Botany.* Botanical Society of America, 1939.
11. LEONARD, J. PAUL. *Developing the Secondary School Curriculum.* New York: Rinehart & Co., 1947. Chap. 7, "Evaluating the School and the Pupil," pp. 209–44; chap. 15, "Evaluating Pupil Learning," pp. 489–512.
12. LEVI, ALBERT W. *General Education in the Social Sciences.* Washington: American Council on Education, 1948.
13. MCBROOM, M. E. *Educational Measurements in the Elementary School.* New York: McGraw-Hill Book Co., 1939.
14. NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. *Forty-fourth Yearbook, Part I: American Education in the Postwar Period.* Chicago: University of Chicago Press, 1945. Especially the chapter "General Techniques of Curriculum Planning."
15. ———. *Forty-sixth Yearbook, Part I: Science Education in American Schools.* Chicago: University of Chicago Press, 1947.
16. ———. *Committee on the Measurement of Understanding,* William A. Brownell, chairman. *Forty-fifth Yearbook, Part I: The Measurement of Understanding.* Chicago: University of Chicago Press, 1946.
17. NEW YORK PUBLIC SCHOOLS, BUREAU OF REFERENCE, RESEARCH AND STATISTICS, Division of Tests and Measurements. *Determining Readiness for Reading.* New York: 1943.
18. ORLEANS, J. S. *Measurement in Education.* New York: T. Nelson & Sons, 1937. Chap. 1, pp. 16–24.

19. PURNELL, RUSSELL T., and DAVIS, ROBERT A. *Directing Learning by Teacher-Made Tests.* Boulder: Extension Division, University of Colorado, 1939.
20. REMMERS, H. H., and GAGE, N. L. *Educational Measurement and Evaluation.* New York: Harper & Bros., 1943.
21. ROSS, C. C. *Measurement in Today's Schools.* 2nd ed. New York: Prentice-Hall, 1947.
22. SMITH, EUGENE R.; TYLER, RALPH W.; and THE EVALUATION STAFF [of the Commission on the Relation of School and College of the Progressive Education Association]. *Appraising and Recording Student Progress.* New York: Harper & Bros., 1942.
23. TABA, HILDA. "The Functions of Evaluation," reprint from *Childhood Education,* journal of the Association for Childhood Education, Feb. 1939.
24. THAYER, V. T.; ZACHRY, CAROLINE B.; and KOTINSKY, RUTH, for the Commission on Secondary School Curriculum. *Reorganizing Secondary Education.* ("Progressive Education Association Publications.") New York: Appleton-Century-Crofts, 1938.
25. TRAXLER, ARTHUR E. *The Nature and Use of Reading Tests.* New York: Educational Records Bureau, 1941.
26. ———. *The Use of Test Results in Diagnosis in the Tool Subjects.* New York: Educational Records Bureau, 1942.
27. TROYER, MAURICE, and PACE, C. ROBERT, for the Commission on Teacher Education. *Evaluation in Teacher Education.* Washington: American Council on Education, 1944.

3. The Functions of Measurement in Counseling

By John G. Darley
University of Minnesota

Gordon V. Anderson
The University of Texas

Collaborators: Margaret Bennett, *Pasadena Public Schools;* A. J. Brumbaugh, *Shimer College;* Galen A. Jones, *U. S. Office of Education;* Robert H. Mathewson, *Harvard University;* E. K. Strong, *Stanford University;* Donald E. Super, *Teachers College, Columbia University*

The Process of Counseling

Counseling is the process in which information about the individual and about his environment is organized and reviewed in such a way as to aid him in reaching workable solutions to a variety of adjustment problems in the normal range of behavior. In this sense counseling is a specialized phase of a personnel service or guidance program. It is important also to realize that counseling covers a normal range of planning and of adjustment problems, albeit this range is rather broad. If we move beyond it in the direction of extreme deviation, we enter the area of medical psychology and psychiatry, involving serious mental illness with or without organic involvement. If we move beyond this normal range in the other direction, we enter the area of routine information-giving and routine orientation, as exemplified in college catalogues, minimum admission requirements, and application procedures. But problems beyond the level of mere information-giving and just short of serious mental illness or social malfunction are persistent, varied, often complex, and often of long standing.

There is, for example, the problem of vocational planning which faces nearly all individuals in our culture. The student, the family, and the educational system will all spend considerable time and effort in arriving at a workable solution to this problem. To a certain extent a definite vocational plan tends to determine the student's educational plans and the particular curriculum he will follow; conversely, the choice of an interesting curriculum not infrequently determines the subsequent choice of a voca-

tion. But even when the problem of curricular choice is seemingly solved by the vocational decision, additional educational problems involving study habits, reading skills, patterns of ability or achievement, discrepancies between achievement and aspiration, and motivation may remain for solution. In the case of a vocationally undecided student, the educational planning itself may become a major task if an effort is made to provide tryout experiences in a variety of vocational fields.

There is also a considerable range of social and personal adjustment problems with which students are faced in the later years of adolescence and the first period of young adulthood. The transition from high school to college often involves new independence, new social relationships, and new social competition. Family relations must be restructured and different social mores must be assimilated.

Since the whole organism is involved in this continuous process of adjustment and readjustment, emotional disturbances, neurotic behavior patterns, and similar maladaptive processes may develop in the individual.

These, then, are some of the problems which students may bring to counselors. For the purposes of this discussion, we are concerned with the problems of students in the later high school years and the college years. Thus, the counselor is here considered a specialist in the diagnosis and treatment of the adjustment problems presented by young people in our highly structured educational system. He is one member of the large group of personal service practitioners that includes doctors, psychiatrists, social workers, visiting teachers, school advisers, and child guidance specialists.

As is true of each of these specialists, the environment in which the counselor works partially conditions his techniques, his activities, and his objectives. Higher education ordinarily involves fairly rigid curricular patterns, for which societal and often legal compulsions to conformity exist. Graduation from an approved medical school is an example of a legal compulsion to conformity before the individual may enter the specified profession, whereas employers' insistence on graduation from a school of journalism or a school of business administration is an example of a societal compulsion to conformity.

Furthermore, those families that send their children on to higher education seem often to view the process as a specialized form of unemployment insurance or social *sine qua non;* it is a necessary finishing process which, in the opinion of parents, can be completed if only their children work hard enough. The fact that approximately six out of ten college entrants fail to survive to graduation gives rise to disappointments and for parents frustrations that may have a serious effect on the personalities of their children.

From the standpoint of the college faculty, education too often involves primarily the accumulation of credits and grades in an orderly sequence; from this accomplishment in this sequence certain inferences are made regarding intellectual growth, maturity, and subject-matter mastery.

The counselor, as a staff member of the educational system, works within the boundaries laid down by these prevailing attitudes, demands, and institutional mores. He may not agree entirely with them; he may recognize their limitations. But this is essentially the institutional framework within which he works, just as certain social workers must deal with the misfits in a highly competitive economic structure, or just as certain doctors must deal with the industrial hygiene problems of a particular industry or area. With respect to the institutional context in which counseling is conducted, it should be remembered that the counseling process itself, and the related activity of measurement, will be affected by the dominating institutional purpose. If the controlling aim is selection, recommendation, or advisement for a particular institutional objective, the procedures of counseling and measurement may be different than if the governing purpose is more general, more concerned with individual growth and development.

By virtue of his training and clinical experience, however, the counselor tends to be actively concerned with more aspects of the whole individual's behavior than does the educational system itself. He has before him always, however dimly seen, a multidimensional criterion of the adjustment of the individual student. He attempts to predict, however accurately, many things to come for each student. He is concerned not only with success as measured by grades earned and requirements completed, but also with success as measured by individual satisfaction and staying power in occupational tasks. He attempts to forecast not only the outcomes of a remedial reading program, but also the effective social adjustment of the individual student, as influenced by social experiences and growth during the college years. He must deal not only with the immediate emotional conflict, but also with speculations regarding the effect of this conflict on the emotional growth of the individual student.

The counselor is essentially a clinician who, as a rule, deals with one student at a time, sees that student against the background of the competitive demands to be met not only in the particular institution but also in the life situations beyond the institution, and employs with that student the most effective diagnosis and therapeutic skills he possesses. Because he carries on this activity in an educational institution, one outcome of his work should be that the student arrives in the classroom situation relatively free from distractions that interfere with learning and more positively oriented toward learning as a means to achieve adult adjustment.

The uses of psychological measurement in counseling, as defined here, are those which relate to individual problems of adjustment, orientation, and development that are brought to the counselors for help, and those which relate to group problems of growth, achievement, educational selection, and classification of students that occur in all institutions of learning. These uses may be summarized as follows:

1. The objective appraisal of personality for better self-understanding and self-direction on the part of the individual himself.
2. The accurate comparison of individual performance with the performance of others for the purposes of selection, recommendation, and self-understanding.
3. Improved basis of prediction as to likelihood of success in any activity in which prospective performance can be measured and compared.
4. Evaluation of personal characteristics in relation to characteristics required for educational and occupational performance.
5. Evaluation of achievement and growth—individual and group.
6. Disclosure of capacity and potentiality as well as the diagnosis of mental disabilities, deficiencies, and aberrations.

The applications of measurement stem from the basic need for objective comparative data upon individual behavior, subject as little as possible to the vagaries of subjective surmise and interpretation.

The Function of Measurement in the Counseling Process

Psychological measurement functions in this counseling process as a means to an end: to help identify individual strengths and weaknesses within an individual and between individuals relative to competitive levels; to provide insight and understanding for the student and the counselor; to structure diagnostic descriptions either in cause-effect terms or in terms of covariation of behavioral elements; and to permit more accurate predictions than would otherwise be possible. One of the continuing emphases of American psychology is to be found in the field of psychometrics broadly conceived. This has affected counseling programs to a great extent. Through the appraisal of human behavior in its measurable elements and the interrelation and predictive power of the resultant data, diagnostic clues are provided for more accurate and meaningful case work.

It should, however, be pointed out that the approach to counseling through psychometrics alone has been misleading, for the counselor's emphasis upon the rationale of measurement and prediction, and the rational nature of the processes through which he gains insight into a counselee's situation, may lead him to lose sight of the fact that what are to him objective data are often highly emotional matters to the counselee. As many

eclectic counselors have long known, effective vocational and educational counseling requires judicious variation of rational, or information-giving, and emotional, or attitude-clarifying, approaches.

Prediction and diagnostic description on the basis of quantitative data are, however, characteristic of many aspects of our social structure. Large-scale industrial and military selection programs can be made more efficient by the development of good predictions of subsequent achievement. Medical science rests heavily on the diagnostic clues and predictive meaning of physiological measurements. Psychiatry and social work profit in part from quantifiable data on individual behavior. In many situations the choice of a particular method of helping an individual to reach a state of satisfactory adjustment is dependent upon the particular problem judged to exist, or the particular diagnosis that has been made.

That some counselors are hesitant to exercise the appraisal function reflects no doubt upon the principle involved, but rather upon the fact that our present knowledge of personality appraisal is rudimentary. This does not mean, however, that we have to give up our attempts to assay and appraise personality either on the ground of individual uniqueness (which no one will contest) or on the ground that ultimate complete knowledge may be impossible. Moreover, even in the present elementary state of knowledge, we may employ such knowledge for what it is worth in counseling situations for the enlightenment of individual clients and for the sake of improving our practices through the discipline of case experience. The individual who receives such information from a counselor may learn something of value about himself not otherwise ascertainable which in no wise would seem to detract from his freedom of personality.

Thus, the "appraisal of the individual" and the conveyance of related personal information is important in the counseling process to the extent that our findings have validity. Meanwhile, we retain this function, even in its present state of imperfection, because it is one of the basic reasons why individuals come to counselors for assistance. Insofar as it can contribute to the improvement of individual and social adjustment, we employ it.

In these respects the appraisal of individuals in the field of counseling may be like that undertaken in the early days of medicine. That developments will occur far beyond what we now know is a foregone conclusion. That individuals will continue to come to counselors for such appraisals is also a foregone conclusion. They will receive direct and explicit help, but this will not necessarily be "determinative" nor "directive."

However, even though advances in psychometric techniques have been of great help to counselors in understanding human behavior, counseling

is considerably more than a matter of measurement alone. The student must be motivated to seek counseling help, in the same way that the patient must want the doctor's help, before either counseling or medical care can be fully effective. And the counselor must be more than a test interpreter. Counselors, and other personnel workers, must recognize at all times that measurement provides only a limited view of certain aspects of personality, the totality of which constitutes a dynamic, unique whole. This means, among other things, that test scores cannot be utilized effectively except in relation to other known, or surmised, factors in the total problem situation. They can never be rightly interpreted except in context.

One of the significant features distinguishing one mode of counseling from another in our current stage of development is the degree of emphasis which the counselor places upon the measured, as against the nonmeasured, aspects of the problem situation as appraised jointly by himself and the client. Some counselors tend to put considerably more weight upon test scores than others.

It is pointed out by those who tend to minimize the value of measurement data that the leeway afforded the individual is sufficiently wide in many actual performance situations and the influence of nonmeasured variables is so high (as, for example, drive or motivation) that the characteristics which we are able to measure may not mean as much in life activity as the nonmeasurable factors.

In recent years Rogers and his students (*18*, pp. 249-52; *19*) have ably presented this viewpoint on counseling in which the psychometric aspects of the total process are minimized. The emphasis is placed heavily on the therapeutic aspects of counseling and on a particular method of therapy. This viewpoint represents a healthy reaction against too great concern with measurement alone. As has been noted in the literature, the psychometric approach to counseling tends to create a passive mental-set in the counselee: expecting tests, he also expects prescriptive answers to his problems. The counselor then has the difficult task of shifting back to the counselee responsibility for analyzing his situation and for making decisions. But if, on the contrary, the counselor begins somewhat nondirectively and uses tests incidentally to the counseling procedure when they seem likely to provide data needed by the counselee, it seems only natural to the counselee that he should himself evaluate the results, checking his interpretations against those of the counselor, and arriving at his own conclusions with the counselor's guidance. This suggests that the use of tests can be reconciled with counseling and with therapy. And it must not be overlooked that some of the debate emerging from Rogers' work relates to differences in nomenclature and methods of therapy, or to the relative importance of diagnostic

descriptions as determiners of the kind of therapy that may be effective. Such differences can be resolved by research and clinical studies. It is quite possible that the need for the kind of precise information supplied by measurement will emerge sooner or later in the majority of counseling relationships, whether the counselor takes the initiative in getting this information or whether the student comes to the point of requesting tests as sources of self-understanding.

Significant Measurements in the Counseling Process

In the design of the ordinary prediction experiment, an attempt is made to use predictors showing relatively low intercorrelations among themselves and relatively high correlations with some criterion of later success. By appropriate statistical treatment, the contribution of each separate predictor can be maximized and weighted into a multiple regression equation that gives the best prediction of the criterion measure. This is essentially an actuarial procedure by which the experimenter hopes to improve, but cannot make perfect, his *selection* for success in the criterion task.

Somewhere in the total process of counseling students it is often necessary to get an answer to a question about the student's chances of success in a given task in curriculum or vocation. To the extent that this student is a member of the sample upon which the multiple regression equation was established, an actuarial, or probability, answer can be given to this question, and the answer derives from test performance. But there are, in addition, factors of maturity, motivation, emotional stability, financial support, and personal adjustment, no one of which is ordinarily itemized in the regression equation and any one of which may determine the success or failure of the individual student. Thus, the counselor finds himself "shading" the actuarial prediction one way or the other, depending upon his assessment of the import of these other factors. The more thoroughly he understands the student, the more conscious he may be of this shading process. No claim is made that the only end of counseling is prediction; this discussion merely emphasizes that prediction is one part of the counseling process, whether it involves prediction of grades, job success, personal adjustment, or marital adjustment. Prediction may be done from rigid application of regression weights or from the exhaustive case study implicit in Allport's (*1*) discussion of personality, wherein so much is learned about one individual that he becomes a unique category not covered by the usual actuarial formula.

For general counseling purposes, a predictive and diagnostic structuring emerges when the following kinds of data are available:

1. General scholastic ability
2. Differential measures of achievement
3. Evidence of special aptitudes or disabilities
4. Interests
5. Personality structure and dynamics, including attitudes and beliefs
6. Socioeconomic and cultural derivation and relations
7. Health and physical attributes

Such items of information meet the general demands of the prediction experiment; each has some relation to subsequent adjustment or success; all are intercorrelated to some degree but not to a high degree; some may act to suppress or depress the effectiveness of other aspects of the individual. Measurement devices of varying degrees of accuracy and validity exist in each of these areas of behavior.

It is, furthermore, a primary thesis of counseling—and more broadly conceived, of psychology itself—that the key to the process of adjustment demanded by life situations in our culture is to be found in the continuous interplay, or balance, or patterning and utilization of these behavioral elements with situational factors. Within broad limits probably the majority of people achieve an integration or adaptive balance that permits orderly adjustments to life demands. In extreme cases maladaptive balances are characteristic and lead to extreme medico-legal solutions. There may be transient periods of disintegration and maladaptation requiring less prolonged forms of therapy or relief. And there may be situations in which the individual needs only a precise item of information about himself or his opportunities in order to move on quickly to a solution of a particular problem.

With regard to the counseling of students, it is illuminating to see the kinds of questions which may be answered to some degree by assessment of the behavioral elements cited above.

1. *Questions regarding vocational planning.* From the limited research data available, it is possible to conceive of families of occupations into which the thousands of payroll titles can be grouped. Jobs within these broad families demand similar patterns of ability, achievement, aptitude, personality, and interests. The measurement of these behavioral elements, therefore, can lead to rough predictions of the job families in which greater or lesser success and satisfaction may be expected. At the college level the choice of a vocation may automatically carry with it the choice of an established curriculum, and the prediction problem may then be reduced to the prediction of success in such a curriculum. Similarly, success and consequent satisfaction in a given curriculum may lead to choice of a related occupation

as the means of continuing the experience of success and satisfaction.

2. *Questions regarding underachievement.* Many students fail to live up to their expectations in college work. The possible causal or conditional factors associated with this problem are numerous and of varying degrees of complexity. At the level of identification (diagnosis), measurement of differential reading skills, of personality characteristics, of differential levels of high school and college achievement may all provide leads or clues to the source of the difficulty and thus point up effective ameliorative or curative opportunities.

3. *Questions regarding personal development and adjustment.* In the normal evolution of a counseling relationship, students are often eager to explore their problems of personality, either with regard to difficult situations or difficult reaction patterns within themselves. Under such conditions test methods of assessing personality structure and dynamics may serve a useful purpose in isolating or defining the area of concern to the student, when combined with appropriate interview techniques in which diagnosis and therapy proceed simultaneously and reciprocally.

4. *Questions regarding motivation and interests.* Even in those cases where differential educational or vocational prediction is clear-cut according to the usual criteria of success, such as grades or earning capacity, the part played by the individual's own interests and desires looms large as a final determinant of his behavior. Again, measurement devices are useful in identifying and classifying some of these comparative interests and motivational factors.

As has been pointed out earlier, the collection and use of psychometric data are only parts of the counseling process. Skillful identification of student problems and effective therapy for them do not follow automatically. But it is important to recognize that counseling, like so many other life situations, involves choice decisions from among various alternative plans of action open to the individual; these choices, in turn, are based on judgments regarding the behavioral factors related to successful outcomes of choice; and these judgments of behavior must be accurately made. Tests are valuable in the extent to which they improve the accuracy of inescapable judgments. The extent to which they do indeed improve the *counselee's* and not just the *counselor's* judgments depends on the readiness of the client to accept the insights of others or the verdicts of experience, and upon the counselor's skill in helping the client to find out what he could more easily, but less effectively, be told. This sequence of events is equally characteristic of the various schools of counseling which appear, in the literature, to represent such divergent viewpoints.

The Research Problems in Human Adjustment

It was suggested earlier that continuing adjustment is a function of the patterning or interplay of various aspects of behavior with situational demands. There is one additional principle that must also be stated: human adjustment follows from orderly processes of growth or development of the personality as a whole and its components. Measurement may reflect or "spot" these orderly growth processes as they unfold, and may indicate deviate or atypical development as well. It is appropriate to specify four properties of measurement devices that enhance their value as aids in making judgments or diagnoses or predictions in the counseling situation, in terms of understanding orderly growth.

There is first the property of *reliability* or *accuracy*. The current methods of test construction maximize the precision and consistency with which behavior can be measured. In counseling situations, as in many other situations, this increased precision is a necessary antidote to tendencies toward over- or under-estimation of relevant behavior by either the counselor or the student.

There is second the property of *validity* or *meaning* or *predictive power*. In spite of all the difficulties of locating adequate criteria of vocational success, or educational success, or personal adjustment, good psychological tests carry some forecasting power for things to come in the life of the individual student, and this predictive value is a balance wheel again in the wishful thinking and perennial optimism with which students approach many long-range decisions.

There is third the property of *economy of effort*. As short, standardized samples of behavior, tests can often supply in a short time and at relatively low cost a basis of judgment-making that is a practical substitute for trial-and-error decisions in our culture. Even if students had endless time and resources to try out a variety of plans aimed at educational, vocational, and personal adjustment, they might still not arrive at a satisfactory end result. Our culture has increased the complexity and variety of choices open to students to such an extent that more economical aids to judgment-making are requisite.

The final property of tests is found in their *normative* aspect, or their indication of an individual's standing relative to others of similar background, experience, or developmental status. This property is related not only to normal growth processes, but also to the extent of deviation from normal growth or status as in certain types of personality measurement. Other chapters of this text will discuss in detail all these properties of tests; they are mentioned here in the context of counseling because they

apply with special cogency to that definition of counseling in which information about the student is so organized as to provide an improved basis for workable solutions to his own problems.

Let us return now to measurement in relation to basic research problems of human adjustment. There are for counseling purposes four fundamental aspects of human adjustment, defined as the interplay of behavior resulting from orderly growth, that must be considered. To each of these, testing has provided both theoretical and clinical insights.

First, the counselor must have some knowledge of the origin, differential rates of growth, and extent of modifiability of human behavior. Without becoming involved here in the heredity versus environment conflict, it is sufficient to point out that much of the information on these topics derives from careful psychometric studies and much of what the counselor does is based either on his interpretation of these researches or on his unsupported and often unverbalized beliefs about the modifiability of human behavior. If, for example, it is believed that hard work is the primary requirement for mastery of any curriculum or any vocation, diagnosis will be aimed at assessing willingness to work hard, and therapy will be aimed at maximizing this willingness. If, on the other hand, it is believed that hierarchies and patterns of abilities are relatively fixed and predictive by the time of late adolescence, diagnosis will be aimed at identifying the pattern most predictive of success, and counseling will consist in part of coordinating this pattern with the student's claimed choices and with the relevant curricular offerings in the institution.

As another example, consider therapy aimed at reducing the disabling effects of a severe emotional state. Such therapy must ultimately take into account the origin and degree of modifiability of the personality structure of the student showing emotional conflict. Therapy is inescapably affected by these factors.

In the more generalized sense, the use of measurement in counseling permits inferences regarding the differential growth points, areas of modifiability, and pattern of organization of the individual student's behavioral elements. Insofar as measurement has contributed to an understanding of this fundamental problem in psychology generally, measurement contributes to its understanding in the case of the individual student.

In the second place, measurement has permitted advances in knowledge of the organization of mental life. Operationally, the relatively low intercorrelations at the college level between measures of ability, aptitude, personality, and interests indicate that these may be viewed as distinct components which must be separately assessed. Within a behavioral element, further analytic breakdowns are possible, illustrated by factor analysis

studies of ability tests, personality tests, and interest tests. Furthermore, the same experimental designs that yield significant results in the study of measured behavioral components can equally well lead to clearer definitions of existing criteria of adjustment in complex life situations associated with educational, vocational, or personal success.

Again, the counselor utilizes this basic knowledge of the organization of mental life in helping the student arrive at workable solutions to his adjustment problems. The issue here is one of choice of tests and other appraisal methods that will provide the clearest and most incisive picture of the significant behavior in the area of ability, achievement, aptitude, personality, or interests.

Human learning is the third fundamental research problem in human adjustment. Progress in construction of achievement tests is related to progress in our understanding of learning. Comparing and contrasting grades and standard achievement test scores as separate indices of learning will provide the counselor and student with significant insights into the educational progress of the individual. The conditions of effective learning, the relation of abilities to amount of learning, the problems of emotional learning and re-education, the motivation toward learning—these are of as much concern to the counselor as to the theoretical psychologist or the educator. In higher education, where learning is especially valued in the institutional structure, many student problems are related to the creation of conditions for effective learning. This situation obligates the counselor to be able to render assistance when and as needed by the students who seek counseling assistance. In this task the use and interpretation of achievement tests are sound and necessary.

The dynamics of personality is the fourth research problem in human adjustment in which measurement has played a role. In today's discussion of objective versus projective, or structured versus unstructured tests, the underlying psychometric properties of reliability, validity, economy, and normative reference still obtain, even though the dynamic emphasis is causing at least a critical review of these properties and their statistical demonstrations. In his day-to-day clinical work, the counselor constantly is concerned with working examples of the research problem. The underachieving student, the student persevering in an occupational choice in spite of failure, the overachieving student, the student afraid to recite in class, the social isolate, or the rejected student, each may pose a problem of personality dynamics to which measurement devices afford clues, insights, or diagnostic definitions that are ultimately helpful in effective therapy.

Essentially, these brief statements of fundamental research aspects of human adjustment, and the psychometric contributions to their solution, tell

something of the equipment the counselor brings to his task. They represent in part the kinds of information which sooner or later must be supplied to the student in need of help so that workable solutions to recognized adjustment problems may emerge.

The counseling interview is the primary vehicle by which this knowledge is organized and transmitted to the student. The counseling interview is also the vehicle in which much of the therapeutic endeavor is contained. While the primary purpose of this chapter is not to discuss therapy and evaluation of counseling, it is essential to touch upon these topics briefly insofar as they relate to the use of measurement in counseling.

Implications for Counseling Practice

A student faced with a particular problem ordinarily must choose a means of attack that will provide some satisfactory solution. For example, in moving from high school to college, the problem of new study habits may occur. First, the student must recognize this new adjustment demand clearly; then he may attempt to build new habits of study to meet the problem. Or when he has to make a curricular and vocational choice, he may ask his friends, follow a family suggestion, estimate his own abilities and interests, and ultimately arrive at a decision. He may seek medical attention for excessive fatigue, weight loss, or eye strain. In each instance, the two processes of identifying the problem and seeking treatment for the problem are characteristic. When the student comes to the counselor and makes his statement of the problem, two people become directly involved in the processes of identification and treatment. The counselor brings to the task not only the measurement skills and fundamental knowledge described earlier, but also some experience in diagnosis and treatment methods. Relatively little is known about the outcomes of specific treatments; the literature is more informative regarding various diagnostic categories. The relation between diagnosis and specific treatment is poorly defined also, but appears to be a significant field for research.

A few examples will illustrate this relationship. When the client is facing a difficult but straightforward problem of vocational choice or planning, he may reach a wiser decision by considering the results of psychological tests and data on previous development and achievement that provide some basis for assessing his potentialities in terms of the requirements of an occupation, the interests and characteristics of those in the occupation, and other related information. At the same time, the tests may furnish information regarding limitations of ability, unexpected assets, and inappropriate interests or attitudes, all of which are relevant to counseling. Such information can then be used by the counselor to assist the client to develop insights

into his own pattern of psychological assets and liabilities as they relate to various occupational and other environmental demands.

It is not to be inferred that possession of clinical data which are yielded by measurement must be followed by prescriptive counseling procedures, in which the course of action which seems wisest to the counselor is mapped out, and the client persuaded to accept it and act upon it. It seems obvious, however, that in the realm of educational and vocational counseling whatever accurate information the counselor can provide the client will be helpful in arriving at a basis for projected action. Skill in counseling involves ability to transmit such information in ways which facilitate its acceptance and use by the counselee.

Admittedly, there are certain client problems which involve primarily a conflict of motives, ambivalence of feelings, or repression of emotional tendencies, which can be greatly aided by wise counseling in which a measurement approach has minimum value from the point of view of therapy. But before this type of therapy is chosen, the application of psychological tests and the review of other clinical data may validly define the area of the client's problem.

Counselors must often make the decision regarding when clients should be referred for psychiatric consultation. When tests of neurotic tendencies and the judgments made by counselors fall in line, it is possible for the counselor to proceed with much more assurance regarding the use of referral in case work. When measurement results are at variance with interview findings, it is usually possible to draw somewhat more extensively on clinical and development data to make a more positive verification.

In the making of diagnoses, counseling problems may be categorized according to various frames of reference. Such categories may have the weakness of pigeonholing clients, but they have the very real value of providing a basis for possible validation studies and studies of the relation between diagnosis and therapy. For example, Bordin (*4*) has suggested that a sociological system of diagnostic classification be replaced by a more purely psychological one. Such an approach is particularly useful to the counselor and student with problems which can be worked out wholly through the counseling relationship. Sociological or situational categories may also serve as the basis for diagnosis, as have been described by Williamson and Darley (*27*) and by Williamson (*23*). Bordin (*4*) has pointed out that in each counseling situation the counselor is faced with a choice of treatment, and discusses this problem in relation to the diagnostic categories used.

Problems which come to the attention of counselors may be those related to the difficulty or inability of the student to cope with a restricted range

of problems in his social environment, or they may be problems representing blocks in the way of student growth and development. Accurate diagnosis should include an understanding, not only of the area of difficulty and the indications of maladjustment (for example, failing grades, unemployment, inability to concentrate), but also of dynamic factors related to the maladjustment (for example, inferior scholastic aptitude, lack of skill or occupational aptitude, repressed sexual strivings, repressed hostilities). It is only through a knowledge of such lacks, conflicts, or inappropriate attitudes that a valid diagnosis can be reached. If the counselor is to assist the client to develop an awareness and acceptance of these dynamic factors, with the expectation that appropriate steps on the part of the client will follow such insights, measurement must aid in the more objective determination of these factors in order to provide a basis for therapeutic approaches and for research in therapy.

Measurement in Evaluation of the Outcomes of Counseling

During the past two decades there have been numerous attempts at evaluating the results of counseling. A general review of methodological approaches which might be used has been made by Williamson and Bordin (*25*; see also *26*), who suggest utilizing the criterion of clinical judgment of adjustment as a basis for evaluating the results of counseling. In numerous earlier studies, college counseling programs have used grade averages in assessing the outcomes of counseling, on the assumption that students who have been well counseled are likely to select programs which are better adapted to their abilities, and in which they will be better motivated; the higher level of adjustment would reflect itself in better work in the curriculum. This is a reasonable assumption, but suffers in its application from the fact that grades in high school and college are quite unreliable indicators of the values which the students may have received from training.

Possibly a more direct, scientific, and reliable approach to the problem of validating counseling is through psychological measurement or appraisal, both before and after counseling, coordinated with the use of appropriate control groups. Such measurement might well include not only tests of educational achievement, but also measures of social and emotional adjustment, and of specific social attitudes. This approach has been urged by Williamson and Bordin (*24*; see also *26*) in the consideration of problems attendant upon the evaluation of individual counseling programs which particularly emphasize educational and vocational guidance and planning; and also by Rogers (*19*) as a means for checking outcomes of "client-centered" counseling in which a nonmeasurement approach is utilized in dealing with problems of a personal nature. Rogers feels that tests should not be used in

therapy, holding that the beneficent effects of counseling are derived from self-insights gained by the client independent of information or interpretations from the counselor, who serves primarily in the role of stimulating and reflecting the client's expression of feelings, and as one on whom such insights can be "tested."

Most studies evaluating the outcomes of client-centered counseling have made use of methods which attempt to quantify interview data, with the inevitable difficulty that a valid criterion must be set aside, while the nature of the counseling process, rather than its effectiveness, is revealed. The investigation of Snyder (*21*; see also *20*) appears promising, however, in that client responses are shown to be possible of categorization in such a way as to yield a criterion, the validity of which few counselors would seriously question. Combs (*6*; see also 7) has demonstrated in an exploratory study the use of a personality questionnaire as a check on the client's progress toward adjustment; in this study, as should be true in all applications of measurement to counseling work, the principle that tests are cross-sectional evaluations which can be used to chart progress is upheld.

It seems likely that in the long run some adaptation of the test-retest experimental design, using control groups and, if possible, discrete diagnostic categories, will appear more frequently as a method of evaluating the outcomes of counseling and of the different therapeutic possibilities subsumed under the general heading of counseling.

Selected References

1. Allport, Gordon W. *Personality: A Psychological Interpretation.* New York: Henry Holt & Co., 1937.
2. Berdie, Ralph F. "Judgments in Counseling," *Educational and Psychological Measurement,* **4**: 35–55, 1944.
3. Bixler, Ray H., and Bixler, Virginia H. "Test Interpretation in Vocational Counseling," *Educational and Psychological Measurement,* **6**: 145–55, 1946.
4. Bordin, E. S. "Diagnosis in Counseling and Psychotherapy," *Educational and Psychological Measurement,* **6**: 169–84, 1946.
5. Bordin, Edward S., and Bixler, Ray H. "Test Selection: A Process of Counseling," *Educational and Psychological Measurement,* **6**: 361–74, 1946.
6. Combs, Arthur W. "Follow-up of a Counseling Case Treated by the Non-Directive Method," *Journal of Clinical Psychology,* **1**: 147-54, April 1945.
7. Cowen, Emory L., and Combs, Arthur W. "Follow-up Study of Thirty-two Cases Treated by Non-Directive Psychotherapy," *Journal of Abnormal and Social Psychology,* **45**: 232–58, April 1950.
8. Darley, John G. *Clinical Aspects and Interpretation of the Strong Vocational Interest Blank.* New York: Psychological Corporation, 1941.
9. ———. *Testing and Counseling in the High School Guidance Program.* Chicago: Science Research Associates, 1943.
10. Dysinger, W. S. "Test Reports in Educational Counseling," *Educational and Psychological Measurement,* **3**: 361–65, 1943.
11. ———. "Two Vocational Diagnoses Compared," *Occupations,* **22**: 304–8, 1944.
12. Erickson, Clifford E. (ed.) *A Basic Text for Guidance Workers.* New York: Prentice-Hall, 1947.

13. FROELICH, CLIFFORD P., and BENSON, ARTHUR L. *Guidance Testing.* Chicago: Science Research Associates, 1948.
14. KITSON, H. D. "Aptitude Testing: Its Contribution to Vocational Guidance," *Occupations,* 12: 60–64, 1934.
15. MUENCH, GEORGE A. *Evaluation of Non-Directive Psychotherapy by Means of the Rorschach and Other Indices.* ("Applied Psychology Monographs," No. 13.) Stanford, Calif.: Stanford University Press, 1947. 163 pp.
16. PATERSON, DONALD G.; SCHNEIDLER, GWENDOLEN G.; and WILLIAMSON, EDMUND G. *Student Guidance Techniques.* New York: McGraw-Hill Book Co., 1938.
17. PEPINSKY, HAROLD B. *The Selection and Use of Diagnostic Categories in Clinical Counseling.* ("Applied Psychology Monographs," No. 15.) Stanford, Calif.: Stanford University Press, 1948. 140 pp.
18. ROGERS, CARL R. *Counseling and Psychotherapy.* Boston: Houghton Mifflin Co., 1942. 450 pp.
19. ———. "Psychometric Tests and Client-Centered Counseling" *Educational and Psychological Measurement,* 6: 139–44, 1946.
20. SNYDER, WILLIAM U. "A Comparison of One Unsuccessful with Four Successful Non-Directively Counseled Cases," *Journal of Consulting Psychology,* 11: 38–42, Jan.-Feb. 1947.
21. ———. "An Investigation of the Nature of Non-Directive Psychotherapy," *Journal of General Psychology,* 33: 193–224, April 1945.
22. STEAD, WILLIAM H.; SHARTLE, CARROLL L.; and OTHERS. *Occupational Counseling Techniques.* New York: American Book Co., 1940.
23. WILLIAMSON, E. G. *How To Counsel Students.* New York: McGraw-Hill Book Co., 1939. 562 pp.
24. WILLIAMSON, E. G., and BORDIN, E. S. "Evaluating Counseling by Means of a Control Group Experiment," *School and Society,* 52: 434–40, Nov. 2, 1940.
25. ———. "Evaluation of Vocational and Educational Counseling: A Critique of the Methodology of Experiments," *Educational and Psychological Measurement,* 1: 5–24, Jan. 1941.
26. ———. "A Statistical Evaluation of Clinical Counseling," *Educational and Psychological Measurement,* 1: 117–32, April 1941.
27. WILLIAMSON, E. G., and DARLEY, J. G. *Student Personnel Work.* New York: McGraw-Hill Book Co., 1937. 313 pp.

4. The Functions of Measurement in Educational Placement

By Henry Chauncey and Norman Frederiksen
Educational Testing Service

Collaborators: Benjamin Bloom, *University of Chicago;* A. B. Crawford, *Yale University;* Henry S. Dyer, *Harvard University;* John M. Stalnaker, *Association of American Medical Colleges*

The increasing importance of measurement in educational placement is attested by the rapid growth of established testing organizations and the founding of new ones, and by the number of local and regional testing programs which are in regular operation. The annual printing of the American Council on Education Psychological Examination alone runs to several hundred thousand copies. The Cooperative tests, measuring achievement in a wide variety of subjects, are being employed extensively by schools and colleges throughout the country. State-wide testing programs, such as those of Minnesota, Iowa, Ohio, and Wisconsin, are assuming ever greater significance in guidance and placement, and in the evaluation of educational programs.

What has caused this rapid growth? It results in part from improved procedures of test handling and from the development of mechanical-scoring devices. The adoption of "objective," or restricted-answer, examinations made it possible to reduce test scoring to a clerical operation, thus considerably lowering costs and increasing the reliability of scoring. The invention of scoring stencils and test-scoring machines has further reduced scoring time and costs, while the employment of Hollerith equipment has facilitated the rapid reporting of test scores and the analysis of data on a scale which would otherwise be well-nigh impossible.

But there is a more basic reason for the new importance of testing: educators have rapidly come to realize that tests offer unique help in many problems of counseling, guidance, and, particularly, of placement. This chapter deals with measurement as applied to educational placement. More specifically, it concerns itself with the use of tests in (1) admission to colleges and to professional schools, (2) articulation, and (3) classification. As the subject of testing in guidance has been covered adequately in the preceding chapter, it will be mentioned here only in brief.

Use of Measurement in Admissions Procedures

College Admissions

Some state-supported colleges and universities are required by law to accept all students who have been graduated from an accredited high school in that state. Mere graduation from high school, however, is insufficient evidence of ability to cope satisfactorily with college work, particularly since the standards of graduation vary from school to school. In such a situation the college must either lower its academic standards or each year fail a considerable number of students. Both these procedures, as well as a compromise solution, have been used in dealing with the problem. At some universities the problem has been met by setting up special courses, sometimes vocational in nature, for those students whose admission is required by law but who are unable to do "college-level" work.

The procedure of eliminating a large proportion of students on the basis of the grades in the first term or half-term, that is, of tryout selection, has obvious advantages. The validity of the procedure is unquestionable insofar as course grades are valid and reliable; correlational studies show a higher relationship between first-term grades and measures of later academic success in college than is usually found between selection test scores and college grades. For example, in a recent study at Princeton a correlation of .78 was found between first-term average and average grade at graduation. On the other hand, the objections to the tryout are not only that it is inefficient and costly, but, even more important, that it may cause a serious sense of failure and frustration in many students.

A university which has a large number of applicants, which is not required by law to admit those students whom it considers substandard, and which wishes to preserve high standards of achievement, is faced with the necessity of devising methods of predicting college success in order to screen out those candidates least likely to succeed. During a period when young men and women in ever-growing numbers are clamoring for admission to college, this problem of selection becomes crucial. College admissions officers are turning more and more to objective tests and other quantitative data for aid in its solution.

In those universities with a selective admission policy, an attempt is usually made to admit only those applicants who are likely to achieve academic success. In addition, there is usually the implication that admission preference is given to those students who possess personal qualities which are likely to make them effective members of the social community.

With the growth of junior colleges, an increasing proportion of students are entering universities at the third-year level. More students are current-

ly being admitted to the University of California as juniors than as freshmen. The problem of admission at the junior level has, however, until recently received relatively little attention.

High school data

The applicant is almost always required to present some evidence regarding his school record. Occasionally the diploma, or evidence that the diploma has been granted, is all that is required. More often, a complete transcript of the school record must be submitted. The use of the transcript varies considerably. The admissions officer may consider it, like the diploma, merely as evidence of graduation. Or, more often, he employs a considerable amount of the data thereon in evaluating the candidate's academic qualifications, or in checking whether special entrance requirements, such as a specified number of units of mathematics, have been met.

The individual subject grades on the transcript may also be used in a more direct way to estimate the applicant's academic promise either for college work in general or for work in specific fields. A generally high record in secondary school is likely to be associated with successful performance in any sort of college program. More specifically, evidence of successful completion of courses in mathematics and science, for example, may be considered as a fair indication that an applicant can handle the program of an engineering school.

Two other types of school data are often used—average school grade and rank-in-class. In general, predictions of college achievement from high school achievement have been found to be fairly accurate; but predictions based on average grade are inferior to predictions from rank-in-class. The reason is that variations occur in standards and in marking systems from school to school, which are uncorrelated either with the ability level of the students or with subsequent measures of success in college.

Rank-in-class is more predictive than average grade because it eliminates some of the variability due to differences in grading practices. It can be used in raw form (for example, 23/181 or twenty-third in a class of 181) or it can be reduced to percentile rank by computing the ratio of standing in class to size of class. If the rank-in-class data are to be used as terminal statistics in the subjective evaluation of applicants, the percentile rank is more convenient and meaningful than the "raw" rank since it permits a direct comparison of students from classes of different size. That is, it assumes that the student who stands at the 90th percentile in a class of 1,000 is approximately equal in achievement to the student who stands at the 90th percentile in a class of 50. If, however, the rank-in-class information is to be used statistically, either in correlation analyses or as one element in a weighted com-

posite of quantitative data, it is necessary to convert the percentile ranks to a standard score on a sigma scale. In many situations it is desirable to report the rank-in-class both as a percentile rank and as a standard score.

Although rank-in-class, however expressed, overcomes the error due to variation in grading practices, it is still susceptible to errors arising from differences among schools in the average quality of instruction and the average caliber of the students taught. A high-ranking student in a school whose pupils have an average IQ of 125 is likely to be more able than a high-ranking student in a school whose pupils have an average IQ of 95. Even more serious, however, may be the errors which result from a lack of uniformity in the procedures used to determine rank-in-class. School A may base the rank-in-class on all students in the school, regardless of the curriculum which they are studying; school B may base the rank only on students in the college preparatory curriculum; school C may compute the rank from the average of the grades obtained in academic subjects alone; while school D may compute it from the averages of the grades obtained in all subjects, including art, shopwork, and physical education. Still another source of error arises from the fact that different schools may use different methods for reporting the rank of students who are tied for the same position. In spite of all these difficulties, however, rank-in-class is usually the best single predictive index available to the college admissions officer. Correlations of the order of .55 are commonly found between rank-in-class and measures of achievement in college.

The predictive value of rank-in-class can be improved somewhat by introducing corrections, based on studies at a particular college, of the achievement of students coming from certain schools. If students from a particular school are found on the average to achieve grades one group higher than the general average, then in the case of a student from that school, a corresponding amount would be added to his predicted grade. At Princeton and Yale, for example, studies have been made of every school from which students come in fairly large numbers, and corrections assigned on the basis of past achievement of students from each school. At Harvard it has been found adequate to correct only on the basis of whether the school was public or private. Further studies of the relative value of the two methods are in progress. Rank-in-class predictions of freshman grades which contain corrections for the specific school or type of school tend to correlate about .60 with actual freshman grades. This correction plan does not, however, help in evaluating the records of students from unfamiliar schools; and in familiar schools it may lead to error when the grading standards shift or the caliber of students changes.

Measures of aptitude and achievement

The prediction of academic success in college on the basis of high school record is fairly successful because of the similarity of the two situations involved. Practically all factors related to academic success in high school—motivation, personal adjustment, study methods, and aptitude—are also operative in college. Prediction on this basis is still far from perfect, however; motivation and quality of adjustment may change, study methods and aptitude which are adequate for high school work may be inadequate for college work, and study techniques may be improved. It is desirable to assess these various factors independently in order to have a more adequate basis than school record alone for admission to college and for individual guidance of the student after he has been admitted. Prediction can frequently be improved somewhat by including with rank-in-class in a multiple regression equation independent measures of aptitude, achievement, or proficiency in tool skills such as reading and writing.

It is impossible to make hard and fast distinctions among tests described as tests of scholastic aptitude, tests of achievement, and tests of proficiency in tool skills. For improving prediction of success in college, all are used for about the same purpose, and there is considerable overlapping in content. A test involving reading comprehension might be published under the title of "reading test" or "scholastic aptitude test" with equal justification. Therefore, in this discussion no rigorous attempt will be made to segregate tests of the various types.

As implied above, most tests used as measures of general scholastic aptitude contain a variety of materials, generally including some verbal and some mathematical items, and possibly items relating to tool skills such as reading comprehension. The justification for calling any test one of aptitude rather than of achievement is usually that the specific items are not definitely related to the content of specific high school courses, although the claim cannot be safely made that the scores on such tests are unaffected by school training. An example of such a test, which is widely used, is the American Council on Education Psychological Examination for College Freshmen. This test yields two scores, a Q (quantitative) score and an L (linguistic) score, as well as a total score based on all the subtests. The Scholastic Aptitude Test of the College Entrance Examination Board also yields two scores, designated as V (verbal) and M (mathematical). Unlike most tests of aptitude and achievement, which are sold to any qualified examiner for his own purposes, the SAT is administered solely by the College Board at its own examination centers, to upward of 70,000 candidates for admission to colleges each year. The Ohio State University

Psychological Test, prepared for the Ohio College Association, differs from the tests mentioned above in that it is entirely verbal and is administered without a time limit.

Many studies have been made of the validity of such tests of scholastic aptitude, and it would be possible to cite page after page of validity coefficients. These coefficients vary from zero to as high as .70. This variability of results is due to many factors, such as the range of talent represented in the group studied, the nature and reliability of the criterion, and the suitability of the particular test to the specific situation studied.

TABLE 2

VALIDITY COEFFICIENTS OF THE C.E.E.B. SCHOLASTIC APTITUDE TEST FOR HARVARD STUDENTS

Type of Student	N	Correlations with Freshman Grades	
		Verbal Score	Mathematical Score
Public school	515	.44	.47
Public school	111	.35	.55
Public school	119	.52	.25
Private school	507	.49	.47
Private school	170	.43	.40
Private school	120	.34	.51

It is probably fair to say, however, that the median correlation of tests such as those mentioned above, using freshman grades as the criterion, would fall somewhere near .45.

Some recent validity coefficients reported by Harvard on the College Board's Scholastic Aptitude Test for several entering groups are shown in Table 2.

In a similar study at Princeton, based on 476 liberal arts students in a recent freshman class, the correlations of the verbal and mathematical scores with first-term freshman average grade were .51 and .30 respectively. For freshman engineering students, the analogous validity coefficients were .42 and .44.

The scholastic aptitude type of test discussed above is ordinarily used to predict over-all success in college. However, in those cases where the test yields both a linguistic and a quantitative score, some *differential* prediction may be possible if the two scores are treated as separate observations of the applicant. Normally, a linguistic (or verbal) score is more highly predictive of success in the social sciences and the humanities than of success in the physical sciences, and the reverse is true of a quantitative score. Where the situation requires more refined differentiation, tests of

achievement in particular subject fields may be needed in addition to the aptitude tests. Such achievement tests may, indeed, serve as important supplements to measures of scholastic aptitude, since they operate on the principle that a reliable and unbiased measure of past performance in a given area provides one of the best means for predicting future performance in the same area.

The Cooperative Test Division of the Educational Testing Service offers a variety of such tests, including not only tests of achievement in specific subject fields at various educational levels, but also tests designed to measure proficiency in the mechanics of writing, reading comprehension, general proficiency in mathematics and foreign languages, and general cultural background. Still other measures of differential aptitude are represented by the Psychological Corporation's Differential Aptitude Tests, the Guilford-Zimmerman Aptitude Survey, the Chicago Tests of Primary Mental Abilities, and the Yale Scholastic Aptitude and Achievement Tests. As a part of its regular entrance examination series, the College Entrance Examination Board also prepares achievement tests in English, social studies, biology, chemistry, physics, mathematics, and the foreign languages. From the ten tests available, the candidate, on the advice of his school or prospective college, may select three.

A number of regional testing programs, usually state-wide, likewise provide the college admissions officer with a wealth of useful data. The New York State Regents' Examinations, for example, provide a method of assessing the achievement of any graduate of a high school in the state, while the Fall Testing Program for Iowa High Schools furnishes scores on a set of general educational development tests. These tests, administered throughout the four years of high school and in the college freshman year, are planned to yield not only measures of proficiency at each particular year, but indications of relative growth or improvement over a period of years as well.

It is difficult to summarize briefly the vast amount of data dealing with correlations between achievement tests and various measures of college success. Considerable variation in predictive value is found, because of such factors as the appropriateness of the test, the range of talent in the group studied, and the method of measuring college success. From a battery of achievement measures we get about as good a prediction of freshman average or other general measure of college success as we would get from rank in high school class; a validity coefficient of about .55 would be typical. Harvard follows the practice of averaging the College Board achievement test scores of its candidates for use in a multiple regression

equation. The zero-order correlations of these averaged scores with freshman grades for several entering groups are shown in Table 3.

TABLE 3

Correlations of Averaged C.E.E.B. Achievement Test Scores with Harvard Freshman Grades

Type of Student	N	r
Public school	515	.55
Public school	111	.55
Public school	119	.56
Private school	507	.54
Private school	120	.60
Private school	170	.54

It should be noted here that the prediction of success in a *particular course* from the appropriate achievement test may be somewhat better than the general prediction; this is particularly true in physical science and mathematics. In these fields it is not unusual to obtain correlations in the neighborhood of .60. Some correlations found at Harvard and Yale are cited in Table 4.

TABLE 4

Correlations of C.E.E.B. Subject-Field Achievement Test Scores with Harvard and Yale Freshman Grades

Test	Course	Harvard		Yale	
		N	r	N	r
C.E.E.B. Mathematics	Freshman mathematics	158	.66	78	.59
C.E.E.B. Chemistry	Freshman chemistry	97	.55	58	.32
C.E.E.B. Physics	Freshman physics	53	.50	28	.58

Those working in the field of predicting scholastic success in college have felt that there are definite limitations to the use of scholastic aptitude and achievement tests. It has been estimated by those who work under conditions as nearly ideal as we can expect that their highest potential predictive value is represented by a coefficient of around .75. And, in fact, even when the best of present achievement and aptitude tests, whose reliability is known to be high, are combined to predict grades, it is seldom possible consistently to attain validity coefficients of more than .70.

Assessment of personal qualities

From the foregoing it is apparent that even when the most valid measures of aptitude and achievement are used there still remains an unpredicted variance in average college grades which amounts to approximately one-half of the total variance. It seems probable that this unpredicted portion is due largely to such factors as persistence, motivation, personal ad-

justment, interest, and study methods—factors difficult to quantify and measure. Probably little further improvement in the prediction of college success can be expected until reasonably valid and reliable measures of such personal qualities have been devised. Such measures are needed not only to improve the prediction of academic success, but also to identify the desirable campus citizen, the individual who can succeed with his fellows as well as in his courses. Perhaps no group of persons is more acutely aware of this need than the admissions officers themselves, and their attempts to meet it have been usually earnest although largely unsuccessful.

The application blank, which frequently contains certain items presumed to yield some insight into personal qualities, is usually thought to be of some value. The applicant may be required to answer questions dealing with extracurricular activities, elective offices he has held in school, educational and vocational objectives, war service, age, and the like. He may even be asked to write a brief autobiography, or a paragraph on why he chose a particular college, or on why he wants to go to college. However, very little work has been done in validating the items of information on the application blank. Presumably a biographical questionnaire could be prepared which would have some validity for predicting college success; the ordinary application blank probably contributes little, if anything, to such a prediction.

Often required as part of the application for admission are letters of recommendation. Such letters are potentially of value, but their very source limits their usefulness. If the applicant is required to provide letters of recommendation, he will undoubtedly apply to friends or acquaintances who will write as glowing accounts of his abilities as possible. Such letters thus have limited value. Letters requested specifically by the university from the school principal are likely to be more useful, although much depends on the person writing the letter. The principal may wish to place as many of his students as possible, and so write uniformly favorable recommendations. On the other hand, he may try conscientiously to write an honest account of each candidate's qualifications. The experienced admissions officer eventually will learn to rely on letters from certain principals and discount recommendations from others. The ordinary letter of recommendation also suffers from the defect of lack of uniformity and systematic coverage of various aspects of ability, achievement, motivation, personal and social adjustment, and the like. In many cases the principal may not be qualified, because of lack of personal knowledge of the candidate, to make such evaluations. Furthermore, the principal's knowledge of the relative excellence of candidates tends to be restricted to the range of talent with which he normally deals. An applicant who appears to him

outstanding in comparison with other students in his school may be mediocre or worse when compared with all the candidates competing for admission to a particular college. Little systematic study has been made of the predictive value of letters of recommendation.

Rating scales have been developed for use at some institutions in standardizing the types of information ordinarily given in letters of recommendation and in standardizing the method of reporting the characteristics of the applicant. It has been found in practice that the ratings usually seem to reflect high school achievement rather than individual estimates of personality characteristics.

Another device frequently used in assessing personal qualities is the interview. The typical interview is unstandardized and of doubtful worth, however, unless its objective is merely to assure that the candidate does not possess obvious defects in appearance, speech, or physique. In connection with college admissions the interview is usually a somewhat hurried and haphazard affair undertaken by an enthusiastic alumnus whose judgment is likely to be clouded by other than academic interests, or by college "representatives" who see the applicants in groups under conditions that do not lend themselves to frank discussion and appraisal. There is some reason to believe, however, that when the interviewer is trained for the job and is permitted to work under good conditions, usable information about the candidate's personal qualifications may be obtained. Recent studies of evaluating officer-like qualities for Army and Navy personnel suggest that reliable and valid judgments of observable personal qualities can be made from a careful interview.

Personality schedules and inventories are rarely used in college admission procedures. Various studies have been made dealing with the predictive value of those personality tests which yield presumptive measures of dominance, submissiveness, neurotic tendency, and the like. The results have usually indicated that such tests bear negligible relationship to college grades, although they may be useful in other ways. Studies dealing with measures of academic interests and of drive or persistence have yielded somewhat more hopeful results, and various other attempts to measure personal qualities have proven to be of some worth. Preliminary studies with the Rorschach test, such as those by Munroe (*11*) at Sarah Lawrence College, seem to indicate that further studies of projective techniques may prove fruitful.

Combinations of measures

The various kinds of data employed by the admissions officer are, of course, used in combination. In a typical situation, the data available for

each applicant might include an application blank, letters of recommendation, a record of impressions from an interview, a score or scores from a scholastic aptitude test, a transcript of the high school record, and rank in high school class. How are all these variables to be used in deciding whether or not the applicants are to be admitted?

A procedure which may be fairly typical in many colleges is simply to sort the folders containing all the available information into groups, based on a subjective evaluation of each individual case. One pile contains the folders of applicants whose qualifications clearly justify admission; another contains folders for all who are clearly too poor to be admitted; and the third contains doubtful cases. The next step is to mull over the doubtful cases and to make additional sortings until enough folders have been added to the "admit" pile to meet the quota for that particular term. Some such procedure is necessary when the data are not quantitative; the sorting is a process of quantifying the data by a ranking or rating procedure.

When part of the data are quantitative, such as rank-in-class and aptitude test score, it is possible to make a definite prediction of the grade of each candidate on the basis of statistical studies determining the validity of these predictors. The validity of the combined measures and their best relative weighting in the predictor can be determined through techniques of multiple correlation. Using multiple correlations for combinations of rank-in-class with various aptitude and achievement test scores, taking freshman grades as the criterion, the coefficient is sometimes as high as .75. The task of making individual predictions on the basis of such studies may be simplified by constructing suitable tables or diagrams. Especially for borderline cases the admissions officer may feel that the prediction can be improved still more by subjectively evaluating interviews, applications, and recommendations. Very little attention has been paid to studying the validity of this subjective type of evaluation or the extent to which such evaluation can improve a prediction made originally from a regression equation. Introducing subjective modifications of multiple regression predictions has been known to weaken rather than to improve the predictions.

Evaluation of admissions procedures

A great deal of study has been devoted to the evaluation of quantitative admissions data—measures of scholastic success in high school, and scores on scholastic aptitude tests, achievement tests, and tests of proficiency in tool skills. A number of universities, such as Harvard, Yale, and Minnesota, have contributed many studies, and the College Entrance Examination Board has long been engaged in studies of its own tests. The value and limitations of these variables have been fairly well demonstrated. One

aspect of the selection problem which has tended to be neglected, however, is evaluation of personal qualities—motivation, persistence, study methods, interests, and personal adjustment. While it is often possible to state with assurance, when a prediction failure occurs, that the student in question failed because of too many social distractions, or because of a lack of objective, or because of emotional problems, it is much more difficult to identify such students *prior* to their admission to college.

The fact that quantitative measures of personal qualities are not readily available is probably the principle reason that little has been done to study the relationships between judgments of such qualities prior to the student's admission and his subsequent college performance. Pending further experiment in the admission situation with such techniques as the controlled interview and projective tests, there are rich opportunities for the investigator who can persuade boards of admission to make an attempt, however crude, to quantify their own judgments of the personal traits of each candidate, as derived from the evidence ordinarily available. The correlation between such judgments and the ultimate success of students, both academically and otherwise, is now unknown. When found for any particular institution, it would almost certainly have an important effect on admission policy.

Evaluation of the assessments of personal qualities obviously requires quantification. One approach might be to attempt to quantify *judgments* of personal qualities without essentially changing the nature of such judgments. For example, all that might be required would be to assign numbers to piles of folders, as described above, and to record the number of the pile for each student's folder. This might be a very crude scale involving only two steps in the continuum, the obviously acceptable and the doubtful (since the unacceptable students would not be admitted and could not be studied); or the folders might be sorted into a larger number of piles representing steps in a continuum. The subjective judgments based on over-all impressions of the data could then be evaluated, using the same correlational techniques ordinarily applied to test scores. Similarly, a study could be set up to compare predictions obtained from a regression equation with predictions obtained after modifying these by applying subjectively evaluated personal data.

Perhaps a more hopeful approach would be to attempt to use techniques of test construction in assessing personal qualities. The possibilities in this area are many. By means of item analysis techniques, using nonacademic college achievement as an external criterion, a personal data blank might be developed which would be considerably more valid than the

ordinary application blank. Or interviewing techniques might be standardized so as to yield some reliable and valid ratings of personal qualities. Further development of tests of persistence or drive, and a follow-up of preliminary studies of Rorschach or other projective tests, again using item analysis techniques with an external criterion, might prove profitable. While some improvement in prediction of college success may come from further refinement of the aptitude and achievement tests, it would seem that the greatest advances may come through a thorough exploration of the measurement of personal qualities.

In any serious research program aimed at improving the prediction of academic success in college, another important first step is to determine the adequacy of the criteria of academic success. What is the reliability of a course grade? What is the reliability of an average of course grades? Which courses tend to be reliably graded and which unreliably graded? What does the grade in a particular course represent in the mind of the instructor who gives it? Is the grade *valid* in terms of the announced objectives of the course? Would more than one grade for each student in a particular course provide a better description of his knowledge and accomplishments? To none of these questions have the experts in measurement yet found satisfactory answers. On dubious evidence, the assertion is frequently and too easily made that course grades, by and large, are unreliable (hence unpredictable) (*a*) because they are too often based on a series of essay examinations whose unreliability is "well known," (*b*) because there is likely to be a large variation in the standards used by different instructors in grading, and (*c*) because grades are often in part determined by such "irrelevant" factors as punctuality in the submission of assigned work, the degree of alertness in class discussions, and the neatness in the preparation of papers. Such notions as these are a part of the folklore of educational measurement. Too little effort has been made to gather the kind of data which would make possible a thorough-going analysis of the situation in any given institution. An important contribution can be made by the expert in measurement who begins his evaluation of admission procedures by first determining the essential characteristics of the criteria they attempt to predict.

Admission to Special College Programs

The problem of college admissions is complicated by the fact that there exist various types of college programs requiring varying patterns of abilities. An applicant might not possess the proficiency in mathematics necessary for successful completion of an engineering course, but be well

qualified for a liberal arts program; or he might be sufficiently handicapped in reading skills to interfere with success in social studies, but do well in a fine arts course.

Successful differentiation of students where admission to special programs must be made prior to freshman year may depend upon opportunities for precollege counseling. The University of Nebraska, which is required by law to admit qualified graduates of accredited high schools in the state, has set up special programs leading to a certificate in two years, for those students who are not qualified for college work at a high level. Before registration, a battery of aptitude and achievement tests is administered to all students, and the test scores, together with information relating to high school achievement and the like, are put in the hands of counselors. Each applicant is then interviewed, and the attempt is made to steer him into the program best suited to his abilities and interests. The success of such a program of precollege counseling depends not only on the counselor, but also on the availability of good tests and dependable descriptions of the relationships between test data and success in the various special college programs. The establishment in 1932 of what is now known as the General College at the University of Minnesota is a more familiar example of an attempt to differentiate among students in admissions to various college programs.

As intimated in the preceding discussion of achievement tests, generally the best predictor of success in a particular subject-matter field, such as physical science or foreign language, is a lower-level achievement test in the same or a corresponding field. The validity coefficients in general compare to those usually obtained between scholastic aptitude test scores and general measures of college achievement. Few studies have been made of possible improvement of such prediction by adding variables, such as high school scholarship in that field, or measures of interest and motivation.

At Harvard, however, a minor study was made to determine the degree to which high school rank-in-class and the C.E.E.B. mathematical aptitude score, when combined with the C.E.E.B. physics achievement score, tended to increase the multiple correlation with grades in a freshman physics course. For one group of 58 cases the physics score alone produced a correlation of .50; the multiple correlation was .62. When a revised physics test was used with a later group of 48 cases, the physics test score alone yielded a correlation of .66, and in combination with rank-in-class and mathematical aptitude score gave a multiple correlation of .69. The difference between the two cases in the amount of improvement is worth noting. An optimally weighted combination of several predictors always improves

the prediction to some extent, but whether the improvement is sufficiently large to warrant the extra labor of applying multiple regression techniques is a matter that must be decided on the basis of evidence pertaining to the particular problem in question.

In certain fields, however, prediction on the basis of aptitude tests has not been successful to the degree that it has in such fields as mathematics, science, and social studies. Comparatively little advance has been made in measuring musical talent, for example, since Seashore's Measures of Musical Talent were developed in 1919. A number of tests dealing with artistic ability have been produced, and some attempt has been made to measure literary discrimination and appreciation. Some of these tests have been subjected to considerable critical study, with varying opinions reported by various authors. It is probably fair to say that tests in this general area are of limited usefulness except when dealing with groups in which a wide range of ability is represented, and that at the college level they are not likely to have great predictive value. Another limitation in differential prediction is in making distinctions within a general field of ability. For example, we are not able at present to distinguish on the basis of aptitude between engineering and other physical sciences. Conversely, there may often be considerable variation in the kinds of talents demanded by different parts of a field grouped under a common label. The prospective chemist who decides to specialize in physical chemistry, for example, will require many skills and aptitudes which are not possessed by his fellow-student concentrating in organic chemistry. Sometimes, but not always, vocational interest tests may be of some help in differentiating among fields which all require nearly the same pattern of aptitudes.

The problem of admitting students to schools of engineering has probably been studied more extensively than that of admission to any other special college program, and the efforts in this area have perhaps been most successful. The types of tests most frequently used, aside from those of general scholastic aptitude, are mathematical achievement or aptitude, spatial relations, and mechanical knowledge or "ingenuity." Since mathematics is an extremely important tool skill in the engineering curriculum, it is not surprising to find that course grades in mathematics or examinations such as the College Board's Comprehensive Mathematics Test are among the best single predictors of success in engineering.

Another aptitude required in many engineering courses, particularly in mechanical drawing and descriptive geometry, is the ability to think in terms of three-dimensional space. A number of so-called spatial relations tests have been developed in the attempt to measure this important ability. The items in such tests are of varied types. They may involve "block-

counting," i.e., counting the blocks in pictures of irregular stacks of blocks, including those blocks not visible but necessary to hold up the structure. Another type is "intersections," in which a plane intersecting a solid geometrical figure is shown; the task is to choose from alternatives presented the figure representing the shape of the resulting cross section of the solid. Still another type requires the student to imagine how a pictured solid figure would look if it were rotated 90 or 180 degrees in one or more directions, while other types of items require judgments on two-dimensional figures. Results on the best of such tests may correlate with grades in mechanical drawing and descriptive geometry as high as .50 to .60, depending upon reliability of the criterion.

There are published a number of "mechanical aptitude" tests. Most of these, such as Stenquist's and O'Rourke's, are heavily weighted with items related to knowledge of tools and parts of mechanical objects. If one supposes that high scores on such a test reflect interest in the realm of mechanics, the test might be assumed to have some value in a differential battery. This supposition is vitiated, however, by the wide variation among students regarding opportunities for earlier acquaintance with practical mechanics. Tests of this type were rather widely used in classifying men for military training in World War II and probably would have some value in predicting success in school or college shopwork, although they were designed primarily for use in industrial selection rather than for colleges. The type of item developed by Bennett for his Mechanical Comprehension Tests was also widely used in Army and Navy classification tests. Bennett's test, sometimes said to measure "barnyard physics," is made up of items which pictorially depict the operation of physical principles in everyday situations. These measures of mechanical aptitude or comprehension have some value in selecting men for vocational training; they have received relatively little attention from engineering schools as aids in selection.

A number of tests of so-called science or engineering aptitude have been developed for use at the college level. Some, such as Zyve's Stanford Scientific Aptitude Test, are inadequately standardized and validated, and probably are of limited value, although certain of the materials seem promising. The subsections of this test are too short to yield reliable measures of the various aspects of aptitude and achievement which they are intended to measure.

The widely used Pre-Engineering Inventory was developed by the Measurement and Guidance Project in Engineering Education, which is sponsored by two engineering groups—the Engineers' Council for Professional Development and the American Society for Engineering Education. The

inventory is a battery of tests designed to measure aptitude and achievement in areas related to success in the engineering curriculum. Composed of six tests, it requires about five and a half hours to administer. The tests measure general vocabulary, technical vocabulary, comprehension of technical passages, tables, and graphs, ability to perform and interpret quantitative operations, comprehension of mechanical principles, and ability in spatial visualizing. The six scores are reported graphically in the form of a test profile showing percentile ranks. Each of the six scores has sufficiently high reliability to justify the use of the individual test profile for guidance purposes as well as for purposes of admissions. An excellent example of what can be accomplished by a serious cooperative research enterprise aimed at solving a particular problem in admissions, the inventory has been found to have good predictive value and has met with approval by engineering schools generally.

At some colleges, limited numbers of students may be admitted to special college programs in the upper-class years, such as the Woodrow Wilson School of Public and International Affairs at Princeton. In such a situation, probably no better method of selection could be employed than to use the academic record of the previous years as the basis for predicting scholastic success. The correlations between sophomore and junior grades in college are ordinarily higher than those between a specific test and junior grades. Little attention has been given to other qualifications that might be required, at least from the standpoint of measurement.

Admission to Graduate Schools

The problem of admitting students to graduate schools is essentially the same as that of college admissions, except that more specific abilities and interests and higher levels of ability are required. The problem is somewhat more difficult from the standpoint of measurement, since a good deal of preliminary selection has already occurred. The range is often restricted still more by requiring that all applicants meet some fairly stringent standard of achievement, such as a B average in college work. Within the comparatively narrow range of ability that remains, it would be surprising indeed if ordinary tests of scholastic aptitude proved to have high predictive value.

As is true at other stages in the academic program, the best predictor of general scholastic success is the record of academic achievement in the preceding scholastic program. Again, however, the accurate assessment of competence from grades is made difficult by variations from college to college in grading standards and practices, quality of instruction, and aptitude of students.

It is now possible to use a battery of achievement tests, known as the Graduate Record Examination, to measure objectively college achievement in eight fields (mathematics, physics, chemistry, biological science, social studies, literature, fine arts, and verbal ability). The battery is suitable for use with college seniors or graduates. Advanced subject tests are also available; ordinarily a candidate would take in addition to the eight general tests one advanced test, probably that in his major field. Now administered by the Educational Testing Service, the Graduate Record Examination was initially prepared through the cooperation of the graduate faculties of Harvard, Yale, Princeton, and Columbia. The examination offers an evaluation of a candidate's general knowledge in various fields as well as his proficiency in a chosen field. Prediction of success in graduate work on the basis of the G.R.E. is about as good as that obtainable from college grades, and prediction based on both G.R.E. scores and college grades is superior to prediction from either one alone. G.R.E. is also valuable to the increasing number of graduate schools which are interested in the broad cultural background of students as well as their aptitude and proficiency in a prospective field of specialization.

A test for graduate students which is similar in form and purpose to the scholastic aptitude type of test discussed above is the C.A.V.D., which has long been used to screen applicants for various graduate schools in Columbia University. The C.A.V.D. is divided into a number of levels, the lowest of which are suitable for use with children in kindergarten and the most advanced of which are appropriate for college graduates. It thus provides a continuous scale for the determination of general scholastic aptitude in both children and adults. At each level, there are four subtests, which give the instrument its name:

C—Completion of sentences by supplying omitted words
A—Arithmetic: items involving the manipulation of numbers and the solving of problems
V—Vocabulary
D—Directions: items which in the upper levels constitute the usual sort of reading comprehension test

The principal difference between the C.A.V.D. and a battery like the G.R.E. is that the former attempts to test the student's ability to reason with material that is not specifically related to courses he may have studied in college.

Studies of the selection of students for programs of graduate study have been confined almost entirely to intellectual aptitudes and abilities, although some attention has been paid to motor skills and interests. Comparatively little attention has been paid by graduate schools to factors of

personality and temperament as they relate to academic or professional success, and little has been done in the measurement of aptitude for most of the highly specific fields of professional specialization, such as patent law, cost accounting, and the various specialties in medicine.

The Medical College Admission Test, successor to the Professional Aptitude Test, is sponsored by the Association of American Medical Colleges and required of candidates for admission by member-colleges. It was first administered in 1948 and consisted at that time of four tests of general ability, including verbal ability on scientific, social, and humanistic materials, and quantitative ability; and two achievement tests in social studies and premedical science. Scores may also be used by approved schools of dentistry and pharmacy as well as by the sponsoring medical schools.

The standards of admission to schools of dentistry are in general somewhat lower than for medical schools, and a somewhat different pattern of ability is required. Skill in many common aspects of the practice of dentistry seems to depend more on motor ability than on high intellectual aptitude. A variety of mechanical and motor ability tests have been tried with varying results, but the prediction of success in dentistry is generally less successful than that in medicine. Tests of intelligence are generally found to be more valid than motor tests, although the relatively low validity of motor tests may be partly a function of the criterion; grading of students on the intellectual aspects of the curriculum is likely to be more reliable and, hence, contribute more to final grades than do ratings of success in manipulating instruments. However, the possible approaches to the problem of measuring motor and perceptual skills have by no means been exhausted.

Widely used for predicting success in law school is the Law School Admission Test, developed in 1947 by a number of leading schools and now administered to candidates four times a year. This test measures verbal ability, including vocabulary, and also reading comprehension, reasoning ability (using fictitious legal cases), and ability to discern crucial relationships. A preliminary validity study shows a correlation of .63 between test scores and first-year law school grades.

Admission to "nonprofessional" graduate schools is typically handled on a more informal and subjective basis than admission to undergraduate colleges. Each department of a graduate school is usually responsible for admitting its own students, and the number of cases therefore tends to be small. The typical procedure probably consists of having the members of a departmental committee evaluate each candidate subjectively, on the basis primarily of his undergraduate transcript, letters of recommendation, and whatever test scores are available, then ranking the candidates in order

of merit. These rankings are later reported at a departmental or committee meeting and the disagreements thrashed out; through some more or less democratic process a final decision is reached on each candidate. The small number of cases usually involved tends to discourage any statistical approach to the problem. In the larger graduate schools, however, it is possible to make studies designed to evaluate and eventually to improve admissions procedures. Such studies are being increasingly performed, and the results tend to justify the continued use and development of measurement techniques in admissions procedures.

Awarding Scholarships

For some time there has been a growing tendency to subsidize the higher education of qualified students, through granting of scholarships, fellowships, and other types of financial aid. The initial impetus for this tendency followed the First World War, when many young people from families of limited means sought admission to colleges and universities. With the onset of the depression in 1929, when financial resources were limited and job opportunities curtailed, many more needy students applied for higher education; this new need for financial aid resulted in the establishment of larger scholarship funds and even, in some instances, state scholarships. Federal aid was granted to many students, beginning in 1934, through the Federal Emergency Relief Administration and later through the National Youth Administration. Following World War II government subsidization of education has been considerably enhanced through the educational provisions of the GI bill, and funds have been provided through the Public Health Service and the Veterans Administration for stipends or part-time work arrangements for training of graduate students in fields where specialists will be needed in the future. An increasing number of scholarships and fellowships is being awarded annually for study at both graduate and undergraduate levels by grants from industrial organizations.

In short, the funds provided for financial aid to students come from varied sources—individual benefactors, philanthropic organizations, religious groups, industry, the state and federal governments, and the educational institutions themselves. The motives of those concerned in the allotment of the funds are even more varied; altruism, income tax exemptions, need for specialists, possibilities of obtaining control over new inventions or discoveries, and the desire for a winning football team have all been involved to a greater or lesser degree. It is not surprising, then, that there is also considerable variation in the procedures for administering the financial grant or award and in the standards required of applicants.

This growing tendency toward the subsidizing of worth-while students

should be considered in its implications for society. The fundamental question to be considered pertains to the ultimate objectives which are to be served by granting financial aid to students. Is it our aim to produce academicians to carry on the humanistic traditions of liberal education? Are we to produce scientists to pursue research considered by our political leaders to be of current importance? Do we wish to train specialists to treat particular social problems? To groom certain young people for political leadership? Or are we to have no definite objective other than to give training consistent with each individual's aptitudes and inclinations in the hope that something of value will ultimately be accomplished? The more specific of these purposes will be criticized on the grounds that they are too narrow and tend to restrict unduly the social and scientific development of society. The last broad objective mentioned above might be considered by some to be so indefinite as to be no policy at all and hence inefficient and wasteful; others may claim that any narrower objective makes for intellectual dictatorship.

Whatever policy is chosen must be implemented by some method of selection as well as by a curriculum and by training techniques. It is with the selection of beneficiaries for financial aid that we are primarily concerned here. Whoever is responsible for such selection methods must face the problem of objectives, whether or not a policy has been determined for him.

University scholarships

In administering a scholarship program at a university, the standards set up are usually analogous to those required for admission except that they are higher. Among other qualifications, the successful candidate must evidence high general scholastic aptitude. This may be determined on the basis of the same tests and indices of school success as are used for admission. Or special scholarship examinations may be required, which are usually similar in form to a scholastic aptitude test. There may also be considered personal requirements, such as high character and qualities of leadership, which are ordinarily judged subjectively on the basis of recommendations and interviews. A third type of requirement is financial need. This may be evaluated on the basis of reports by students, usually endorsed or supplemented by financial statements from parents. The student may be required to furnish an itemized budget showing all sources of income and financial resources. Improved methods of estimating the fair share of expenses which should be borne by the student's family are being developed at some universities. Since scholarship awards are generally made from year to year, the recipient, in order to retain his scholarship, is

usually required to maintain a specified standard of achievement in terms of grades. Where funds are made available by donors, certain additional restrictions may exist, such as geographical region, race, religion, field of study, school attended, or even family name.

The awarding of university scholarships is not always an objective and impartial matter. And since techniques of evaluation are most subjective in the area of character and personality, it is here that abuses are most likely to occur. Alumni groups may, for example, throw strong support to a candidate on grounds other than those of scholastic achievement and aptitude, with results which are occasionally unfair to those who lack influential partisans. The application of carefully developed measurement techniques, where available, tends to put the awarding of scholarships on a fairer basis. The improvement of techniques for awarding scholarships involves essentially the same problems of measurement as have already been discussed in connection with college admissions.

Methods used by other organizations sponsoring scholarships

Increasingly scholarships are being sponsored by organizations other than universities—business and industrial organizations, philanthropic foundations, and religious groups. The usual arrangement is for the funds to be administered by a university or by an administrative group composed primarily of educators, presumably to avoid any semblance of an outside control of higher education.

Likewise, the federal government has avoided control of education. A veteran of World War II must be admitted to a college through its regular admission procedures before GI funds are made available for tuition and subsistence, and no restrictions are set up by the government as to course of study. In the new NROTC program there are again no restrictions as to course of study beyond the hours which must be devoted to naval science, and each NROTC college is free to exclude any student, even though he has been accepted by the Navy through its own selection program.

In recent years a number of scholarship programs sponsored by non-educational groups have been set up on a nation-wide basis, with awards based on competitive examinations. The Westinghouse Electric and Manufacturing Company has, for example, established a number of scholarships for engineering training. Under the name of "Science Talent Search," a scientific aptitude examination is administered annually to a large number of high school seniors. This aptitude test is made up of items involving number series, vocabulary, mechanical movements, mathematical movements, mathematical reasoning, and spatial relations, arranged in omnibus fashion, together with a section on science reading comprehension. Each

competitor qualifying on the aptitude test submits an essay on some scientific project he has himself carried through. A small group deemed to have produced the best essays are then brought together for several days in Washington, D.C., with a group of recognized scientists, who thus have ample opportunity for appraising the personal qualifications of each candidate. The final awards are made at the close of this period on the basis of all the data available. The interesting feature of this method of selection is that it applies the so-called "successive hurdles" approach, thus supplying very complete information on which to judge the intellectual and personal qualifications of the "finalists" from among whom the winners are selected. Its principal disadvantage is one common to all selection methods based on successive hurdles: highly desirable candidates may be eliminated in the early rounds of competition because of insufficient evidence. While the method unquestionably identifies excellent candidates, it may fail to identify all the *best* candidates.

The same general method is employed by the U.S. Navy in the selection of candidates for its college programs. Under the terms of the Holloway Plan, the U.S. Navy is admitting to the naval and naval aviation college program more than two thousand young men each year under conditions equivalent to generous scholarship awards. These men are admitted to the program through procedures set up by the Navy; they must also be accepted by the NROTC college or university through its regular admission procedure. The first step in the Navy's selective process is the administration, at special centers throughout the country, of a selection test containing items of the scholastic aptitude type. Through this examination a preliminary screening is made. The candidates are previously informed about certain other requirements, such as freedom from certain physical defects, and those candidates who know of factors which would prevent their admission to the program are discouraged from applying.

Those who survive the preliminary screening are directed to report to an Office of Naval Officer Procurement for further processing, which consists of a physical examination and interviews. Only those candidates who survive the physical examination are interviewed. Considerable care has been taken in designing a standardized interview situation, in the course of which interviewers use standard rating forms; each candidate is interviewed independently by two officers trained in the technique of interviewing. All the information—test scores, interviewers' ratings, and a composite score—then goes to a regional board, which determines which of the candidates are to be accepted. All these procedures are definitely oriented toward fairly specific objectives—the selection of men capable of successful college work and endowed with qualities which will make them

successful officers in the United States Navy. The successful candidates enjoy midshipman status and its equivalent in pay while they are students. This whole program of selection has been planned with unusual care and may serve as a valuable proving ground for testing certain selection procedures.

Scholarships and fellowships for graduate study

The awarding of fellowships and scholarships for graduate study at most graduate schools is usually a part of the admissions procedure. As described previously, the typical methods are largely subjective evaluations of such materials as transcripts and letters of recommendation, culminating in a rank-ordering of the candidates. The best fellowship might be offered to the highest-ranking candidate, and so on down the list. Fellowships established by industrial or business organizations are usually administered by the graduate school in the same manner as the university's fellowships.

The National Research Council and the Social Science Research Council are awarding predoctoral fellowships which they, rather than the university, administer; the methods of selection are generally similar to those employed by the universities themselves. The National Research Council uses, in addition to other evidence, the results of a battery of aptitude and special field tests.

While the foregoing is only a brief review of methods of administering student scholarships and other financial aids, it may give some notion of the variability in objectives and techniques in current usage. The objectives are both specific and varied; yet we do not find a corresponding variation in the measurement devices used, except to some extent in the content of instruments for appraising aptitude. Verbal, quantitative, spatial, and mechanical comprehension items are used in varying proportions, depending upon the objectives of the scholarship program. We do not, however, find much similar variability in appraising personal qualities. With our inadequate knowledge of how personality characteristics are related to success in different fields of academic endeavor, we cannot now say whether or not distinctions can be made in this area. Perhaps the desirable personal qualities are the same, regardless of the field of specialization; or perhaps research will eventually make it possible to designate differentially and measure crucial personal qualities.

Use of Measurement in Articulation

In the educational progress of a student through the elementary grades, high school, and college, it is, of course, extremely important that he be enrolled in courses which are appropriate to his level of proficiency at any

given time—that he take courses which are neither too difficult nor which involve wasteful duplication of earlier-learned content. This problem is particularly acute at the times of transition from elementary school to high school and, even more so, from high school to college, because at these steps there is ordinarily a break in the continuity of supervision and a change in administrative control. The problem of articulation has to do with techniques and procedures for bridging these transitional points so that there is continuity in the educational program and so that each student is enrolled in courses which are consistent with his interests and level of academic proficiency.

Various factors tend to interfere with good articulation between elementary and high school and between high school and college. The wide variability in standards of achievement and in level of talent from school to school, for example, tends to handicap good articulation; a certificate of graduation at a certain educational level is no guarantee that any fixed standard of proficiency has been met. Another factor is the increasing flexibility of curricular programs; course requirements tend to be less rigorous than formerly, and in some schools superior students may be encouraged to advance far ahead of their classmates. Finally, some school administrators fail to realize the problems of articulation and consequently make small effort to solve them.

That the problem of articulation is real and far from solution is shown by several studies made by educators. Such studies show, for example, that a large proportion of the content of some freshman college courses is a duplication of high school courses; that a fair proportion of high school seniors may surpass the median of college sophomores or even seniors in certain achievement test scores; and that superior high school students are frequently able to show good proficiency before taking certain elementary college subjects. It has been shown that superior students who were admitted to college at the end of the third year of secondary school did as well as students with four years of such schooling. These and many other studies indicate that there is often considerable inefficiency and wasted effort in the educational program, particularly for the superior group. No less serious is the converse of this problem; many students are placed in high school or college courses for which they lack the basic preparation. The growing tendency to institute special remedial courses in mathematics and reading at the college level is symptomatic of this difficulty.

There are at least two important aspects to the problem of articulation. One pertains primarily to curriculum planning, to the proper organization and coordination of the courses of study at the school levels involved.

This topic has been discussed in chapters 1 and 2. The other aspect, pertaining to the proper placement of each individual student, is the major concern of the present section.

Use of Tests in Improving Articulation

First of all, the improvement of articulation between both elementary and secondary school programs and between secondary school and college programs depends upon adequate guidance facilities. Since a counselor or faculty adviser cannot perform his duties effectively without adequate information about not only the requirements of the various courses but also the aptitude, level of proficiency, and interests of each student, he should have this latter information available to him through two sources: educational records and tests, both general and differential.

Within school systems where there is coordination of administrative control, the problem can be handled fairly well through educational records, provided that grades in specific subjects are based on reliable tests of achievement. The Educational Records Bureau has contributed to the solution of this problem by cooperating with schools and school systems in planning record systems and conducting testing programs.

Frequently, however, records furnish an inadequate solution because schools have varying standards, curriculums, methods of instruction, and grading systems. This is particularly true in articulation between secondary school and college, where there is seldom any continuity in administration, and students may come from schools widely scattered throughout a state or even the nation. In such a situation the use of special tests of aptitude and achievement is particularly useful in improving articulation.

Articulation of secondary school and college programs

The tests used in improving articulation between secondary school and college may be administered either near the time of graduation from school or after admission to college; the first method has the advantage of making the test data available for use in admissions procedures as well as in articulation.

The Regents Examinations in New York State, originated in 1877, are the oldest state-wide achievement testing program in existence. Examinations are given to high school seniors throughout the state in a wide variety of subjects ranging from English, American history, and French, to music, art, typewriting, and shorthand. Scores on these tests when reported to the college of the student's choice furnish valuable data for the admissions officer and also for advisers who are concerned with the problem of specific course assignments. Careful use of such information can help to avoid the

many academic difficulties caused by placing students in courses for which they are either over- or under-prepared.

The aptitude and achievement examinations of the College Entrance Examination Board are also administered near the end of the senior year in high school and thus provide colleges with information which can be used to articulate the college freshman program with the high school courses of incoming students. To make these or any other such tests maximally useful, each college employing them must conduct local studies to establish standards for admission to particular courses. Depending on the nature of the course, such a study may take the form either of predicting what the scores mean in terms of subsequent success or of simply establishing norms on known populations. At Harvard, for example, it has been found appropriate to determine course placement in freshman mathematics by predicting the student's probable success from his C.E.E.B. mathematical aptitude score. In various language courses, however, definite test norms have been established by administering the tests to students who are on the point of actually completing the courses at different levels. Course placement in this latter situation is achieved by comparing the test performance of the incoming students with the various norms so established.

The distinction between these two approaches to the problem may be important. In the case represented by mathematics, the background of all incoming students is roughly comparable, so that the aptitude score is, in a general way, a measure of the facility with which the student can handle quantitative and symbolic concepts. In a sense it provides a measure of the *slope* of the student's learning curve on the subject, and success in freshman mathematics appears to be predictable from a knowledge of this slope *regardless of the actual position on the curve*. In other words, the decision is between a "slow" course and a "fast" one. In the case of the languages, however, it appears that the best placement is made from a knowledge of the student's *position* on the learning curve. The decision depends on whether the student has learned enough in secondary school to enable him to cope with the content of a particular course.

Various other regional high school testing programs are in progress, such as those of Iowa, Minnesota, and Illinois, and the scores are being put to use by both schools and colleges within the state.

One objection to such testing programs, whether on a state-wide or national scale, is that a college adviser cannot count on each of his group having taken the tests which are appropriate to his individual problems. The student may have escaped the test because he came from a region where such tests were not used or, in the case of the C.E.E.B. examinations, he may have made an inappropriate selection. These annoying irregularities

can largely be overcome by setting up a freshman testing program at the college. If such a program is carefully planned and tests rapidly scored and reported on, pertinent information for practically every freshman can be in the hands of each adviser prior to the time of registration. Such a program, however, obviously can serve only to improve articulation, and cannot help in admissions work.

The Cooperative Test Division of the Educational Testing Service, previously mentioned, provides a large number of achievement tests which are appropriate for placement purposes, and in its annual National College Freshman Testing Program recommends a battery of tests to be administered to the entire class in an institution. Also available for this purpose are the Iowa Placement Examinations, which have been rather widely used.

The University of Chicago has developed its own set of College Placement Tests, prepared by its Board of Examiners. These tests cover various fields such as American history and government, humanities, physical sciences, mathematics, English, and foreign languages. The scores obtained by an individual on various pertinent tests are used in assigning him to particular courses. The tests are divided into Common and Special placement tests, of which all students take the former. If the student's performance is sufficiently high, he is excused from certain first course requirements, and students with unusually high scores are invited to the Special placement tests. If the student's performance on the Special test is sufficiently high, he is excused from certain advanced courses. Theoretically, it is possible for a student to meet all the requirements of the College of the University of Chicago through the passing of placement tests. The division of the tests into Common and Special insures that only very able students will need to take more than twelve hours of placement tests.

Articulation of elementary and secondary school programs

While the problems of articulation at this level are essentially the same as those between high school and college, it is often administratively easier to provide the basic achievement test data here, because elementary schools and high schools are usually operated as part of the same public school system. It is thus easier to organize a unified procedure of achievement testing and record-keeping and to see that adequate elementary school records accompany each student to his secondary school.

A number of commercially published achievement tests are available at this level. Among the best known and most widely used may be mentioned the Iowa Every-Pupil Tests of Basic Skills, the Metropolitan Achievement Test, Modern School Achievement Tests, the Progressive Achievement Test, and the Stanford Achievement Test.

Articulation within the school

The problem of articulation is, of course, not confined to the period of transition from one school to another; it is also essential that, as the student progresses from grade to grade within a school, he be enrolled in courses which are as well fitted as possible to his particular level of proficiency and to his educational-vocational objectives. The problems here, at least for the elementary school, largely involve the relation between the curriculum and the maturity of the children.

Initially the child's educational level is established for the most part by his age. As he progresses up the educational ladder, it may become apparent that he is placed in a grade which is either too advanced or too retarded for his development. Some children are intellectually qualified to work at a more advanced level than most of their classmates; others, at a lower level. Readjustments may be made as a consequence of such situations. The usual expedients are nonpromotion and "skipping" (double-promotion) or acceleration. The limitations of these procedures are presented in chapter 1.

The problem is, of course, not as simple as is implied in the foregoing paragraph. Factors other than intellectual maturity, such as social and physical development, must be taken into account, and the psychological repercussions of such a drastic step as failure to promote must be weighed. Even "intellectual" maturation may be uneven, so that a pupil is qualified for more advanced work in some respects only; for example, he may be exceptionally proficient in arithmetic but retarded in reading. Procedures for providing for individual needs in heterogeneous groups, as treated in chapter 1, should probably receive greater emphasis in the schools. Suitable tests can make important contributions to the solution of such problems. The next topic, "Use of Measurement in Classification," deals with some particular aspects of the problems involved in attempting to place pupils in the educational environment which is best adapted to their degree of growth and development.

Use of Measurement in Classification

In recognition of the wide individual differences among pupils in learning ability, adequacy of preparation, interests, industry, social maturity, and so on, the attempt is often made to segregate pupils into more or less homogeneous groups so that group methods of teaching may be effective. The simplest of these attempts places pupils in separate school grades solely on the basis of age. Perfect homogeneity is impossible to attain and, as a matter of fact, would probably be undesirable; but some degree of homo-

geneity beyond mere age classification is generally agreed to be desirable. A teacher finds it impossible in a classroom situation simultaneously to interest and encourage the bright students and to prod the dullards up to some minimum standard of achievement. The attempt to place school children in homogeneous groups is referred to as "classification" or, sometimes, "sectioning" (especially in the upper grades). The purpose of classification is to reduce the range of individual differences in instructional groups and thus make possible a type of instruction and educational environment more nearly suited to the needs of each individual student. Since classification has already been discussed in chapter 1 in reference to the improvement of the learning situation, it is unnecessary to discuss further its advantages and limitations for this purpose.

Classification has been based on various types of criteria, used singly or in various combinations. The most appropriate criteria are those which give the best prediction of achievement. The types of measures in common use (other than age) include records of achievement in previous grades, tests of general mental ability, achievement tests, and reading readiness tests.

The type of measure to be used depends somewhat on the organization of the school. If pupils are to be classified into groups which remain the same for all parts of the curriculum, general measures are appropriate; while if pupils are classified differently for different courses, such as reading and arithmetic, then more specific types of measures are indicated.

The predictive value of tests of general mental ability is fairly high for average achievement and achievement in subjects such as reading and arithmetic. They are far less successful in predicting proficiency in spelling and handwriting. Achievement tests and school grades are successful in about the same degree as mental ability tests in predicting later achievement. These general statements are based on a large number of separate studies, the results of which are by no means uniform. It would be necessary for each school to study the predictive value of various measures and to use those which are most successful with its own methods of instruction and techniques of appraising achievement.

Reading readiness tests are widely used in the early grades to determine whether the child's perceptual processes are sufficiently mature to enable him to make the fine visual discriminations that are necessary in the transition from speech to written symbols. The Gates Reading Readiness Tests, which are among the most widely used, require the child to perform such tasks as finding which two of four printed words are the same, finding which one of four printed words is the same as the one shown on a card by the examiner, naming letters and numbers, and finding the picture in a

set the name of which sounds almost like the name of another picture. Such tests tend to correlate about .50 with mental age. Reading readiness tests have been found to be the best predictors of achievement in reading and have been used as one basis for determining when the teaching of reading should be begun for the individual child.

Even with the most careful attempts to bring together for instruction students who are homogeneous with respect to their ability to learn a particular kind of material, a good deal of variability is bound to be present. Individuals are such complex creatures that, even with the best measures so far created, it will not be possible to discover students who are similar in *all* aspects of motivation, interests, skills, and scholastic aptitude. On the other hand, by proper use of suitable tests it is possible to decrease significantly the variability with respect to particular abilities and skills that are found to be important in the teaching of a certain subject.

It is, of course, possible to classify pupils on the basis of their achievement during a trial period of, say, three to four weeks. The advantage of tests is that the trial period may be eliminated, thus avoiding the feelings of insecurity which frequent changes of class are likely to produce in children, and permitting a quicker adjustment of both teacher and pupils to the classroom situation.

This chapter has discussed problems related to admissions, articulation, and classification for which testing techniques have already offered some partial solutions, and has indicated other problems where equal success may be encountered in years to come. In all such applications, the limitations of tests and of measuring devices, as well as their potentialities, must be considered. As research on educational measurement is continued by universities and testing organizations, better solutions to the problems yet unsolved will doubtless be forthcoming. In the meantime, the importance of an experimental approach by each educational institution to its own problems of selection and placement by whatever means are available cannot be overemphasized. All educational problems cannot be settled by the large research organizations; the methods of making optimum use of educational and psychological measurements must ultimately be discovered by each school in the light of its own particular situation.

Selected References

1. BENGTSON, NELS A. *Annual Report of the University Junior Division.* Lincoln: University of Nebraska, 1941.
2. BUROS, OSCAR K. (ed.). *The Third Mental Measurements Yearbook.* New Brunswick, N.J.: Rutgers University Press, 1949.

3. Crawford, Albert B., and Burnham, Paul S. *Forecasting College Achievement.* New Haven: Yale University Press, 1946.
4. Donahue, Wilma T.; Coombs, Clyde H.; and Travers, Robert M. W. *Measurement of Student Adjustment and Achievement.* Ann Arbor: University of Michigan Press, 1949.
5. Dyer, Henry S. "Evidence on the Validity of the Armed Forces Institute Tests of General Educational Development (College Level)," *Educational and Psychological Measurement,* **5**: 321–33, 1945.
6. Findley, Warren G. "A Group Testing Program for the Modern School," *Educational and Psychological Measurement,* **5**: 173–79, 1945.
7. Frederiksen, Norman. "Predicting Mathematics Grades of Veteran and Non-Veteran Students," *Educational and Psychological Measurement,* **9**: 73–88, 1949.
8. Garrett, Alfred B. "Giving College Credit in Chemistry by Examination," *Ohio Schools,* **24**: 356–57, 1946.
9. Learned, William S. *Examinations and Education.* Carnegie Foundation for the Advancement of Teaching, Forty-First Annual Report, 1945–1946. New York: The Foundation, 1946.
10. Learned, William S., and Wood, Ben D. *The Student and His Knowledge.* ("Carnegie Foundation for the Advancement of Teaching Bulletin," No. 29.) New York: The Foundation, 1938.
11. Munroe, Ruth Learned. *Prediction of the Adjustment and Academic Performance of College Students by a Modification of the Rorschach Method.* ("Applied Psychology Monographs," No. 7.) Stanford, Calif.: Stanford University Press, 1945.
12. Ross, Clay C. *Measurement in Today's Schools.* 2nd ed. New York: Prentice-Hall, 1947.
13. Ryans, David G. "A Study of the Observed Relationship between Persistence Test Results, Intelligence Indices, and Academic Success," *Journal of Educational Psychology,* **29**: 573–80, 1938.
14. Schrader, William B., and Conrad, Herbert S. "Tests and Measurements," *Review of Educational Research,* **18**: 448–68, 1948.
15. Segel, David. *Prediction of Success in College.* ("U.S. Office of Education Bulletin," 1934, No. 15.)
16. Traxler, Arthur E. *Techniques of Guidance.* New York: Harper & Bros., 1945.
17. Wood, Ben D., and Haefner, Ralph. *Measuring and Guiding Individual Growth.* New York: Silver Burdett Co., 1948.

Part Two

THE CONSTRUCTION OF ACHIEVEMENT TESTS

5. Preliminary Considerations in Objective Test Construction

By E. F. LINDQUIST
State University of Iowa

COLLABORATORS: Henry Chauncey, *Educational Testing Service;* Walter W. Cook, *University of Minnesota;* Warren G. Findley, *Educational Testing Service;* T. R. McConnell, *University of Minnesota;* Ralph W. Tyler, *University of Chicago;* Ben D. Wood, *Columbia University*

FOR THE PURPOSES OF THIS VOLUME, THE CONSTRUCTION OF AN educational achievement test will be regarded as consisting of five major steps, as follows:

1. Planning the test
2. Writing the test exercises
3. Trying out the test in preliminary form and assembling the finished test after tryout
4. Determining the procedures and preparing the manuals for administering and scoring the test
5. Reproducing the test and accessory materials

Each of these steps will be separately considered, so far as the objective type of test is concerned, in chapters 6–11. Problems of test construction relatively unique to *performance* and *essay* tests will be considered in chapters 12 and 13 respectively. The scaling of the test and the establishment of norms, which, in a sense, are also a part of the task of test construction, will be discussed in Part Three, "Measurement Theory."

This analysis of the steps involved in test construction takes for granted that certain preliminary decisions have already been made concerning the general nature and purposes of the test to be constructed. Preliminary to actual test construction, the author has presumably selected, or specified in broad general terms, the educational objectives the attainment of which is to be measured by the test. Presumably, he has also specified, at least in general terms, the population of individuals with which the test is intended to be employed, the conditions under which the test will be administered and used, and the major uses to which the test results are intended to be put. For example, the test author may already have decided that his test

is to measure the ability of high school students to use the English language correctly in their own writing, and that the test is intended to serve primarily as a basis for educational guidance of high school students and for the evaluation of high school instruction in this area. Again, the test constructor[1] may have decided that his test is to determine how well the students in a particular course in United States history at the ninth-grade level have attained the immediate objectives of instruction in this course, with the intention that the test will be used as the principal basis for assigning course grades and determining whether the student has passed or failed the course.

It should be apparent that the decisions which are made preliminary to actual test construction are, from the broadest point of view, far more important or crucial than those which follow. The contributions of educational measurement to the educational process as a whole, and to its improvement, depend as much or more upon the ability of test constructors to recognize the situations in which tests are most needed or can do the most good, or upon their ability to anticipate the good and bad effects of the projected tests upon educational practices, as they do upon the technical competence of the test constructors. It is, of course, important that whatever tests are constructed measure validly and dependably whatever they do attempt to measure, but it is even more important that what they do attempt to measure be worth while and significant, and that through their use the tests and test results exercise a desirable influence upon the aims, habits, attitudes, and achievements of students, teachers, counselors, and school administrators.

The major purpose of this chapter is to review and evaluate the broad trend of these preliminary decisions in the recent history of educational achievement testing, and to consider the possible need for a fundamental redirection of these preliminary decisions in the future. A second purpose of the chapter is to consider briefly the various possible "approaches" which the test constructor may take to the measurement of educational achievement in general. Upon considering the difficulties of measuring the objective that he has tentatively selected, the test constructor may often decide to forego its measurement entirely, and may turn instead to other more easily tested outcomes. The second purpose is thus very closely related to the first. The chapter as a whole, then, is basically concerned with the problem of deciding *what* to measure, but serves as well to introduce

[1] For convenience in this discussion, the term "test author" or "test constructor" will be used in place of the more appropriate term "test construction agency." In modern practice, the "author" of a standardized test is usually a team of individuals, including specialists in subject matter as well as in the techniques of test construction.

the problem of *how* to measure, with which the remaining chapters of Part Two are concerned.

The Selection of Objectives

Ultimate vs. Immediate Objectives

Many of the basic objectives of school instruction cannot possibly be fully realized until long after the instruction has been concluded. For guidance in specific courses of instruction, however, it is common practice to set up less remote objectives—objectives which are capable of immediate attainment. Ideally, these immediate objectives should in every instance have been clearly and logically derived from accepted ultimate objectives, in full consideration of all relevant characteristics of the pupils who are to receive the instruction. Ideally, also, the immediate objectives should be supported by dependable empirical evidence that their attainment will eventually lead to or make possible the realization of the ultimate objectives. Finally, the content and methods of instruction should, ideally, be logically selected, devised, and used with specific reference to these immediate and ultimate objectives, and should likewise be supported by convincing experimental evidence of their validity.

Unfortunately, this ideal relationship among ultimate objectives, immediate objectives, and the content and methods of instruction has only rarely been approximated in actual practice. Some of the content of current instruction, if derived at all from sound and accepted ultimate objectives, has been derived from them by a process of faulty inference, and contributes much less to the realization of the objectives than other content which could be substituted for it. More unfortunately, a portion of the present content of school instruction is there only by reason of the organization of the curriculum by "subjects," and because of the practice of introducing new materials in intact subject units, or subject by subject, often without any careful selection of the detailed content of those subjects. As a result of this practice many detailed elements which have no relationship whatever to any ultimate objectives have entered the curriculum simply because they "belonged" in the same broad category of knowledge, or in the same subject, with other content which could be readily justified, and because of which the subject as a whole was selected. A considerable portion of the factual content of current instruction in courses in United States history, for example, is of this character. This is not to say that the place of this subject in the curriculum cannot be as readily justified as any other. Knowledge of certain facts in American history undoubtedly contributes to the

realization of certain important educational objectives. It does not follow from this, however, that *any* historical information has a place in the curriculum. It is not the teaching of United States history per se which contributes to the objectives, but the appropriate use in instruction of certain specific historical materials which have been carefully selected with these objectives in mind. Nevertheless, many historical facts are included in the school textbooks for apparently no better reason than that they are of interest to historians or that they are a part of "American history," and many otherwise useful facts are taught in such a way as to minimize their contribution to desirable educational objectives. History courses are not, of course, alone in this respect. Similar statements, with more or less justification, may be made about every subject in the school curriculum.

As a result of this organization by subjects, furthermore, many other elements of content may survive in the curriculum long after their usefulness has gone, simply because they "belong" to a subject which is still, on the whole, of unquestioned value, or which retains its place in the curriculum, in spite of its questionable value, by reason of tradition, inertia, and the power of vested interests.

With reference to subjects of the type last noted, it is frequently true that, instead of the content and immediate objectives of instruction having been derived from any accepted ultimate objectives, just the reverse of this relationship obtains. That is, many of the immediate objectives were actually derived from, or adapted to, the traditional content; and their claimed relationships to ultimate objectives—as well, sometimes, as the ultimate objectives themselves—were "thought up" in an effort to rationalize the continued teaching of the traditional content. It has been claimed, for example, that knowledge of Latin will contribute to improved reading comprehension in the vernacular, through developing a better understanding of English words and phrases of Latin origin, or through developing a more accurate grammatical sense, and so forth. While there may be some truth in these claims, they are clearly made in an effort to justify the perpetuation of Latin in the school curriculum, rather than to justify its selection in preference to all known alternative and more direct ways of improving English reading comprehension. Again, of course, Latin has been used only as a convenient example. For any subject in the school curriculum, particularly for the long-established subjects, many of the immediate objectives claimed for them have a similar origin. Indeed, one of the most common of all *real* objectives in teaching, in general, is "to teach the facts contained in the textbook." Such objectives as "to acquire a sound knowledge of world geography," "to know the important facts of United States history," "to understand common natural phenomena," are often only an-

other way of saying, "to know what is in the text." The really functional objectives of many school subjects—the day-by-day objectives that most teachers are actually trying to attain—are, in large part, *content* objectives of this type.

The foregoing is not intended as a general indictment of the practice of organizing the curriculum by subjects. The practice clearly has its advantages as well as disadvantages, and many of its undesirable features could be eliminated while retaining the advantages. Whatever one does about curriculum organization, there will always be the problem of deriving immediate objectives from those more remote or fundamental and of allocating objectives to appropriate units of instruction. Furthermore, while "subjects" may remain the same in name, they may be, and frequently are, changed in content and method so that certain ultimate objectives may be more effectively realized. Our concern in this discussion, then, is only with the implication of the present subject organizations and content of the curriculum for achievement testing, and not with the pros and cons of subject organization in general.

Immediate Objectives of Traditional Subjects in Achievement Testing

The situation just reviewed is one of particularly serious concern with reference to educational achievement testing. Practically all of the standardized achievement tests and informal school examinations that have been constructed to date have been designed to measure achievement in established school subjects. At the high school and college levels, particularly, standardized achievement tests that are intended to measure the attainment of general objectives, or that disregard or cut across subject-matter boundaries, have constituted only a very small proportion of the total offering. Furthermore, these tests, almost without exception, have been based on the actually functioning (as opposed to the claimed or theoretical) immediate objectives of instruction—often with little regard to the possible lack of relationship of these immediate objectives to any important ultimate objectives of the entire educational program. Where no authoritative statements of immediate objectives have been available, or where such objectives have not been sufficiently specific or meaningful, the test constructor has often set up or derived his own immediate objectives for the test. He has usually derived these objectives, however, not from any general ultimate objectives, but from the common content of current instruction. Analysis of textbooks and courses of study has constituted the most common technique for validating educational achievement tests. Accordingly, just as the real objective of instruction has been "to teach what is in the textbook," so,

in many instances, the real objective in testing has, to an even greater degree, been "to test for recall of what is in the textbooks," or "to test what is now being taught."

Even where the prevailing immediate objectives of instruction have been most seriously questioned, test builders have continued, for purposes of test construction, to take these immediate objectives for granted. Achievement tests for high school Latin, for example, have been concerned exclusively with the students' ability to read or write Latin, and their builders have made no pretense whatever, in these tests, of measuring the extent to which this ability contributes to the realization of any of the basic objectives of the whole program of general education. Tests in high school mathematics similarly have measured such things as the ability to factor polynomials, with little regard to the probable social utility of such skills, etc., etc.

On the whole, then, efforts of test builders to improve educational achievement tests have been largely confined within the framework of the prevailing subject-matter organization of the curriculum and of the generally accepted immediate objectives of instruction. Within these restrictions a vast amount of progress has been made, but it has consisted primarily of the introduction of further technical improvements and refinements, providing for increased comparability of scores from test to test, for increased reliability and efficiency in measurement, for better coverage or sampling of the content of instruction, for more representative and more reliable norms, and for improved administrative and scoring procedures. The improvements that have thus been made, however, have been real improvements only in relation to certain relatively narrow classroom uses of tests (which will be discussed later). The greater opportunity, that of providing better tests for the purposes of individual student guidance and of curriculum evaluation in relation to a more enlightened concept of general educational objectives, has been seriously neglected.

As has just been implied, achievement test builders have not been without considerable justification for concerning themselves so exclusively with the construction of subject tests based on prevailing immediate objectives. Most achievement test builders will readily grant the truth of the preceding comments, at least so far as the typical school is concerned, and are quick to recognize the need for tests and other instruments that may be used in curriculum evaluation and in guidance. Many of them would contend, however, that a test cannot be highly valid for purposes of curriculum evaluation and at the same time also be appropriate for most of the uses to which achievement tests are generally put. That is, many of them would

assert that there is a need for two distinct types of tests. While recognizing the need for evaluation instruments, they have felt that the construction of such instruments involves taking a responsibility for curriculum modification that they, as test technicians, are unwilling to assume. In other words, many test builders have taken the position that, while it is their responsibility to keep pace in their tests with accomplished curriculum changes, it is not their business to *bring about* changes in the curriculum. Their position, further, would be that as long as large numbers of teachers are teaching a subject in accordance with certain generally accepted immediate objectives, those teachers have a right to know how well they have accomplished those objectives—no matter how far those objectives may be out of line with the most advanced thinking concerning more remote objectives, or how little the subject as a whole really deserves its present place in the curriculum.

Most test users—that is, most teachers of special subjects—would readily subscribe to the latter point of view. Teachers of special subjects are not, in general, held personally responsible for the place of those subjects in the curriculum of their school, nor even, in most instances, for the immediate objectives and detailed content of those subjects. They are given certain subjects to teach, and often the content is prescribed for them in detail (that is, the textbooks have been selected for them). They are, however, held personally responsible for "getting the content across" to their pupils, and all too often the results of subject tests have been used to hold them thus responsible. Accordingly, the more exclusively the tests used are concerned with that which the teachers have actually been trying to accomplish, and the less the tests are concerned with things for which the teachers do not feel responsible, the more acceptable the tests are to them.

Most teachers, for these reasons, prefer standardized tests that differ from their own informal examinations only in that they are provided with norms, and that they are more reliable, more objective, more easily scored, and more highly refined technically, than their own product. In his own informal examining, the typical teacher has favored questions that can be passed only by students who have taken his particular course, and who are directly and intimately acquainted with its unique organization and content. Questions that can be answered on the basis of general information, or which many of the students could have answered before taking the course, or to which they could have learned the answers outside of class, are for those reasons suspect and are excluded from the examination. Whether consciously or unconsciously, many teachers, in their own informal testing, have tried to build examinations that will discriminate as sharply as possible between students who have taken and those who have not taken their

particular courses. The result has often been, for example, that two equally good students taking courses of the same title but under different instructors (even in the same school) might each fail on the examination intended for the other student, even though each might make an A on the examination intended for his own course. The standardized tests which are most closely adapted to the local curriculum in this sense are those most often selected and used by the subject teacher.

Further justification for the practice of basing achievement tests only on prevailing immediate objectives may be readily adduced from the point of view of the individual student. Standardized tests are widely used in the assignment of school marks, in the measurement of school progress, for purposes of pupil motivation, and in the maintenance of standards. With reference to these uses it may be argued that the pupil is not responsible for instructional objectives, and that if he accomplishes well what he has been asked to accomplish he should be given credit and rewarded for it; that it would be unfair to penalize him for failing to learn something which he had neither been encouraged nor given an opportunity to learn; and that the use of the same test for purposes of curriculum evaluation and pupil motivation would only confuse the pupil rather than motivate him to greater effort.

For the reasons given, tests of course achievement concerned primarily with current instructional objectives will continue to be constructed and used, as long as the subject organization of the curriculum is retained. Attention will later be given in this chapter (pages 138–40) to the question of how and by whom such tests may best be constructed. The only point to be made here is that such tests should no longer occupy so dominant a place in the whole educational scene as they have in the past, nor should they absorb nearly so much of the thought and effort of agencies constructing tests for wide-scale use. The arguments presented in the preceding paragraphs do *not* constitute an adequate general justification for past practices in the construction and use of educational achievement examinations. As a result of these practices, educational achievement tests and achievement testing in general have exhibited a number of very serious limitations, particularly in relation to guidance and evaluation, and have been incapable of measuring or of assuring adequate recognition of many of the most important aspects of the total educational development of the pupils. It will be shown in the following sections how these limitations have been due, in part, to the emphasis on content and on questionable immediate objectives, and, in part, to the strict adherence to the subject-matter organization of the curriculum. Consideration will then be given to the undesirable consequences, with reference both to the improvement of instruc-

tion and of educational guidance, of the almost exclusive use of tests with these limitations.

Limitations Due to the Emphasis on Content Objectives

There is nothing objectionable about content objectives, as such, in instruction, so long as the content is used in a manner appropriate to the outcomes ultimately sought, and so long as its proper relation to the ultimate objectives is recognized and understood. There can, of course, be no instruction without specific instructional content, and one of the most immediate and clearly legitimate objectives of instruction is to help the student make the best use of this content. This does not mean that much of the content should be memorized for purposes of later recall. Much, if not most, of the detailed informational content of instruction in the social studies, for example, is presented to the student with little or no expectation that he will be able to recall it in accurate detail long after the course of instruction has been concluded. The skillful teacher uses this content in order to develop broad and meaningful generalizations, to establish trends, to illustrate principles, to bring out relationships, to raise problems, to exemplify and explain procedures, to characterize institutions, to develop attitudes, to establish desirable habits and ways of thinking, to develop the ability to evaluate evidence, to develop a good sense of values and sound judgment, and so on. The teacher rightly demands *temporary* learning of some of this detailed content, but primarily in order that the content may be more effectively used for the aforementioned purposes during the course of instruction. It is the generalized outcomes which are expected to be permanent; it matters relatively little if the student soon forgets many of the *detailed* facts from which the generalizations were originally derived, or by means of which the generalized procedures, skills, habits, attitudes, and ways of thinking were originally developed. As a result of an effective course of instruction in United States history, for instance, an adult may have retained permanently a very adequate picture of, say, Colonial America, highly accurate in broad outline and in all essential particulars, and quite adequate as a basis for understanding contemporary American institutions, traditions, and ideals, and yet he may have forgotten nearly all of what he once knew in the way of exact and detailed information about specific events and personalities, about specific wars and campaigns, explorations and discoveries, congresses and conventions, legislative acts and reprisals, and so forth. He may, similarly, have acquired a good understanding of the evolution of our present system of transportation, expressed in generalities quite adequate for purposes of thinking about contemporary problems, and yet may have forgotten entirely such facts

as the time of completion, the cost, and the statistics on use of the Erie Canal. More important, even though he has forgotten such detailed facts, he may have retained, with little loss, whatever contribution the study of these facts made to the enrichment of his vocabulary, to the establishment of desirable attitudes toward democratic institutions, to his ability to do critical thinking about contemporary social, economic, and political problems, etc. To a considerable extent, then, memorization of detailed content is only a temporary and incidental outcome of instruction—a means to an end, rather than an end in itself.

Closely related to the need for only temporary learning of much of the detailed content of the social studies is the wide variation permissible in the selection of the specific content used to attain the same generalized outcomes of instruction. Two instructors in the same subject, for example, may both wish to develop in their students a generalized understanding of the nature and purposes of American labor unions. Obviously, a very wide choice of illustrative materials is available for this purpose—far more is available than need be employed or can be employed in a single course of instruction. One instructor may thus decide to use one set of specific examples, while the other may use an entirely different selection of instructional materials, yet both may serve essentially the same ends. On a somewhat higher level, different instructors may concern themselves with quite different ideas, generalizations, problems, and so forth, yet accomplish essentially the same purposes so far as still more remote objectives are concerned. One instructor, for example, may devote considerable attention to international trade relationships, a topic which the other may neglect in order to give adequate consideration to domestic financial problems and policies. One may stress certain national cultures and ideologies; another may select still others for intensive consideration. Yet the two courses may, for all general educational purposes, be completely interchangeable. Even different subjects may be regarded as interchangeable with reference to still more remote objectives, as is implied and recognized in the practice of allowing students to elect different high school and college courses in the same broad areas in meeting general educational requirements.

Comments similar to the preceding may be made about any broad field of subject matter, particularly at the level of general education. Various instructors in one of the natural sciences, for example, may all strive to develop in their students an understanding of the role of science in modern society, of the techniques and methods used by the scientists, and of the limitations as well as the possibilities of these methods in answering questions about the world. All may be interested in building up in their students the basic scientific vocabulary needed in reading and thinking

about contemporary scientific developments, and all may wish to improve the ability of their students to observe scientific phenomena and to draw valid inferences from their observations. Yet, different instructors might use widely differing instructional content with equal success. As in the social studies, even different subjects may prove equally effective in relation to such objectives. A course in physics may serve such purposes as well as would one in chemistry; a course in botany as well as would one in zoology. Again, the fact that the student is unable, months or years afterwards, to recall much of the detailed content of a course of instruction in science fortunately does not imply that the instruction was futile or that the course held no permanent values for him.

In the field of literature, likewise, one instructor may acquaint his students with certain outstanding literary works, while the other may employ an entirely different selection of literary materials. Both sets of instructional materials, however, may be used with equal effectiveness for the purpose of developing in the students a better understanding of their fellow-beings, of their motives, ideals, and frailties. Both may be equally effective in developing the ability to read literature with comprehension and enjoyment, or in developing improved tastes in the selection of reading materials, or in establishing desirable and lasting reading habits. There is probably no other field of instruction that permits wider latitude in the choice of instructional materials, or in which such freedom of choice is more desirable, than the field of literature.

The foregoing is not meant to imply that the content type of examination has no validity whatever, or that there is no legitimate place in educational achievement testing for tests or test items that hold the student responsible only for the recall or recognition of specific items of information. The case for the continued use of tests of this type rests primarily on two considerations. In the first place, some of the content of instruction in any subject *is* taught with the hope that it will be permanently retained by the student. No item of information, of course, is taught purely for its own sake—every bit of specific content in the course of study must ultimately be defended in terms of the number of uses that will later be made of it, either during, or subsequent to, formal instruction, and in terms of the importance or cruciality of these uses. Some content is useful only in accomplishing certain immediate purposes in instruction, as in establishing a single generalization or in illustrating the application of a single principle. Such purposes, furthermore, may frequently be served equally well by many other similar items of information. Some content, on the other hand, is so widely and so obviously useful in adult living that, if it is retained at all, its later utilization may safely be taken for granted. Some of this con-

tent, furthermore, may be the only content that will serve these particular purposes. With such materials, one of the teacher's primary concerns should be with the learning, for indefinite retention, of the content itself. With such materials, also, the so-called content examination, granting that it does not test for *rote* memory only, is quite appropriate whether used during the course of instruction or at any subsequent time. There is, on this basis, a very definite place in educational achievement testing, both in subject examinations and in tests of general educational development, for items of the informational type. In general, however, only a minor proportion of the total content of instruction is of this character, particularly in the subjects making up the program of general education at the high school and college levels. Only a relatively small proportion of the content items in the typical subject examination may be defended on these grounds.

The case for the content examination rests, in the second place, upon the belief that even with content which is primarily of immediate instructional value only, there is some positive correlation between the student's ability to recall this content and the extent to which he has derived any generalized values from it and will retain them permanently. Facts which have been well organized in the student's mind will probably be better remembered than those which are unorganized or unrelated. The formulation and retention of a generalization probably helps the student remember the facts used to establish it. Because of this incidental relationship, a content examination may always have some validity as an indirect measure of the attainment of certain ultimate objectives. The relationship, however, has never been shown to be very high. Few instructors would be willing to have the effectiveness of their own instruction judged, on an absolute basis, by the performance of their students even on their own content examinations, if these examinations were "sprung" on their students several years after they had taken their courses, nor would many instructors care to have their students so judged, even on a relative basis, as to the extent to which they had profited individually from these courses.

All fundamental objectives of education are ultimately concerned with the modification of behavior. The extent to which any individual is genuinely educated depends not upon what he knows nor upon the amount of information that he has acquired, but upon what he is able to *do*. For the reasons just noted, there is, nevertheless, some positive correlation between the extent of the student's knowledge and his true educational status. It is not surprising, in view of the ease and reliability with which the extent of the student's knowledge may be determined, that the earliest efforts to objectify the measurement of educational achievement relied so strongly

on this relationship—that is, that the first objective tests were almost exclusively of the content type. This approach to the measurement of educational achievement, or, particularly, of the attainment of ultimate objectives, has already been exploited to the full. If it represented the only available approach, the limits of improvability of educational achievement tests would already have been reached. Fortunately, this is not the case. The opportunities for further improvement are unlimited, but they lie in the direction of more direct measurement of ultimate and general objectives, rather than of more reliable measurement of immediate and content objectives.

One very important implication of these limitations of the content type of examination remains to be considered. It follows, both from the interchangeability and the temporary value of most of the content of instruction, that the typical subject examination is particularly inappropriate or lacking in validity when used with individuals who not only have not recently taken a course of the same title, but who have never taken any such course. That is, the typical subject examination is most inappropriate when used with individuals whose educational development in the particular area covered by the test has been acquired from informal out-of-school experiences or incidentally through the study of other subjects, or when used with individuals who are largely or wholly self-educated.

It was the recognition of these limitations of available subject examinations which led, in part, to the construction of the United States Armed Forces Institute's Tests of General Educational Development. The authors of these tests were given the responsibility of selecting or constructing a battery of tests that could be used to determine the general educational status, or the appropriate placement in a program of general education, of war veterans returning to school at the close of World War II. It was apparent that many of these veterans would return to school much farther advanced educationally than when they had left school to enter the service—that while their *formal* education had been interrupted by the war, their *education* had continued, although perhaps usually at a considerably slower pace than had they remained in school. It was even more apparent that whatever educational development had taken place while they were in service would have come about in a markedly different fashion than if they had remained in school. Their in-service education would have been obtained, not through textbooks or classroom instruction, but through direct observation and experience enriched through travel, through the reading of newspapers, magazines, and books, through informal discussions and consideration of personal and group problems, as an incidental to specific military training, and through systematic study, either organized for them or self-initiated and self-directed. From these experiences, the veterans derived

many of the same generalized values—of the nature reviewed in the preceding paragraphs—that might otherwise have been acquired through formal school instruction, but quite obviously most of the "content" from which these values were derived not only differed markedly, in general, from the traditional content of formal school subjects, but also varied greatly from one individual to another.

It was suggested early to the authors of the GED tests that the educational status of the returning veteran might be satisfactorily determined through a carefully selected battery of available standardized tests corresponding to the principal subjects in the typical curriculum, and that the veteran might be given "credit" for each subject in which his test performance equaled that of civilian students who had actually taken and passed the course in question. Very little deliberation was required for the rejection of this suggestion. It was clearly evident, for the reasons reviewed in the preceding paragraphs, that the use of even the best of available subject examinations would fail to reveal the veteran's true educational status, and would penalize him severely because of the manner in which his education had been acquired. Many veterans, for instance, who had been forced to leave high school before graduation had, nevertheless, through their in-service educational experiences, acquired the practical equivalent of a general high school education, and should on that basis have been granted a high school diploma or equivalency certificate. To have required those veterans to satisfy the usual school standards on a battery of subject examinations would have been just as unfair as to have required all veterans who had actually graduated from high school to pass such examinations again upon leaving the service in order to be permitted to retain their high school diplomas. What was needed, obviously, was a battery of tests which would provide far more direct measures of the ultimate outcomes of a general education than could be secured through any available subject examinations.

Few educators will deny the shortcomings of the typical subject-matter examination in the situation just reviewed. Yet, if such examinations are inappropriate for measuring the general educational status of war veterans, they must necessarily be inappropriate for measuring the general educational status of pupils who are yet in school. Opportunities for informal or incidental learning or for self-education are by no means open only to those who are no longer in school. On the contrary, a very considerable share of the educational development of pupils at any level of education may be attributed to out-of-class experiences. In the case of many of the more intelligent, intellectually curious, observant, and well-read youngsters, particularly those with favorable home and community environments, op-

portunities for summer travel, and rich job experiences, and so forth, it is quite possible that the *major* share of their total educational growth has involved experiences outside of the classroom. With such individuals, as much as with war veterans or with self-educated adults in general, any collection of content or subject-matter examinations is certain to fall far short of revealing their true educational status.

In concluding this description of the limitations of the content examination, it may be well to note that much of what has just been said about these limitations would still apply, even though those examinations had been based entirely upon the best possible content, that is, upon content which, for purposes of achieving the desired ultimate outcomes in specific courses of instruction, was as well selected and as appropriate to these purposes as any content could possibly be. In consideration, therefore, of the fact that much of the content on which these examinations had typically been based in the past was not thus carefully selected, but was included in the curriculum and in the examination primarily as a result of tradition and historical accident, the limitations of examinations based upon the prevailing content objectives become all the more evident.

The preceding discussion has been concerned primarily with the limitations of subject examinations of the content type, that is, tests designed to measure the extent of the student's *knowledge* of particular school subjects. Not all immediate objectives of instruction, of course, are content objectives, nor are all subject examinations of the content type. Of the remaining immediate objectives, most are concerned with the development of specific skills and abilities. The situation in educational achievement testing with reference to objectives of the latter type is, in general, much more satisfactory than in the so-called content subjects. This is true, in part, because in these areas of instruction there is frequently a much more direct and obvious relationship than in the content subjects between the immediate objectives of instruction and the ultimate objectives of the whole program of education. In many cases, indeed, there is no clear distinction to be made between immediate and ultimate objectives, that is, the ultimate objectives are themselves immediately attainable. Many of the things which the pupil is taught to do in elementary school arithmetic, for example, are exactly the things it is hoped he will later be able to do as an adult. Thus, tests of computational skill and of problem-solving ability in arithmetic, if appropriate for use during the course of instruction in arithmetic, should be equally appropriate for the measurement of those skills at any later time or in any other situation. A skill may deteriorate after it has once been taught, of course, either because of an inadequate maintenance program in

subsequent instruction, or because it finds no application in life outside the school, but the validity of the *test* does not deteriorate subsequent to instruction, as is true of many content examinations (see page 152).

The foregoing is not meant to suggest, of course, that all immediate instructional objectives of the skills or abilities types now generally taught may be readily defended, nor that all skills are taught which should be included in the program of general education. There are many high schools, for example, in which all pupils are required to learn how to divide higher order polynomials and to factor quadratic equations, as well as many schools in which pupils are given no specific instruction in such obviously useful skills as those required in map-reading or in the interpretation of statistical tables. Whatever the nature of the immediate objectives of instruction, the fact that they are currently prevalent in teaching is no safe indication that they constitute a valid basis for the construction either of subject examinations or of tests of general educational development.

Limitations Due to the Subject Organization of the Curriculum

The modern American school accepts responsibility, in some degree, for every major phase of the child's development—intellectual, emotional, social, moral, and physical. It is particularly responsible for all aspects of his intellectual development. Each and every teacher in the school, whatever his specific assignment, shares in this comprehensive responsibility. The teacher of physics, for example, is not only a teacher of physics, but is also to some degree a teacher of reading, of correctness in speech and writing, of mathematics, of the social studies, of the fine arts, and so forth. At the same time, he acts in the role of friend and counselor to the pupil, and advises him or influences his decisions on many educational, vocational, and personal problems. His job is not merely to teach physics, but to help each individual boy and girl become a better-educated, more effective, and better-adjusted member of society.

It is, accordingly, extremely important that each teacher acquire a very comprehensive acquaintance with each child under his care, that he be at all times dependably informed concerning all major aspects of the child's development. There is, thus, a very definite and continuous need, in instruction, in educational and vocational guidance, and in curriculum evaluation, for *comprehensive* descriptions of the total educational development of individual pupils and of groups of pupils—descriptions based as much as possible upon dependable and objective measurement. Because of the subject organization of the curriculum, however, practically the only objective measures on which teachers, counselors, and administrators have in the past

been able to base any descriptions of the pupils' educational status have been scores on *subject* examinations, for practically no other type of tests of educational achievement (with the exception of reading tests) have been constructed and used, particularly at the high school and college levels. Strenuous efforts have been made, especially by guidance workers, to make the most of the test information thus collected, particularly through the maintenance and use of individual cumulative records of test scores and related data. Thus far, however, these efforts have met with only indifferent or limited success. The descriptions of individual pupil development which it has been possible to derive from these test records have, at best, been sadly lacking in comprehensiveness, while the measures obtained have been extremely difficult to interpret because of lack of comparability, and have been incapable of yielding satisfactory indications of educational growth.

The descriptions have been lacking in comprehensiveness, in the first place, for the very obvious reason that measures of an individual's educational status at any given time have been restricted to the areas of instruction in which he happened, at that time, to be taking courses. No measure of a student's mathematical ability, for example, has been obtained at all unless he chanced, at that time, to be enrolled in a course of mathematics; no measures of his development in the social studies were obtained in the years when he was not taking a course in that area, etc. However, even though all such restrictions upon the administration of subject examinations were removed, so that at the time of graduation, for example, a student might be given a subject examination in each of the subjects he took while in high school, the resulting description would still be seriously incomplete and unbalanced. This would be true, even apart from the factors noted in the preceding section, because of the fact that the recognized ultimate objectives of instruction of individual subjects do not collectively constitute or account for the recognized ultimate objectives of the whole program of general education. Satisfactory adjustment in family and marital relations, for example, is generally recognized as an important objective of any program of general education at the high school level. Yet this objective would rarely be found listed among the recognized and *functioning* objectives of any individual high school or college subject, and scant attention, if any, is given to it in any subject examination. Any general objective of this character—for which no particular course has been offered, or for which no particular course or series of courses has been made especially responsible—tends to be neglected both in instruction and in measurement. In general, complex problems demanding the integrated use of knowledges and skills acquired in several courses or areas of instruction practically never appear in any subject examinations; the student's abilities to deal with such

problems are neither measured by the examinations nor adequately developed in formal instruction. Incidentally, these are important added reasons why batteries of subject examinations were judged inappropriate for measuring the total educational development of war veterans (page 132), or of informally educated or self-educated people in general.

Past achievement test records (scores on subject examinations) have provided only very inadequate indications of individual pupil development, furthermore, because of the lack of comparability in the measures obtained. Extended consideration will be given later (chapter 17) to the problem of rendering comparable measures that have been secured from different tests. It may be noted, in advance of this description, that the techniques now available for making scores comparable depend on the possibility of administering the different tests to the same or to comparable groups of individuals. Standardized examinations designed for special school subjects, however, have in nearly every case been standardized only for populations of pupils currently taking the subject for which the test was designed. This means, of course, that with very few exceptions no two subject examinations have been standardized for the same or for demonstrably comparable groups of individuals. Thus, a percentile norm for an algebra test, based upon a sample of ninth-grade pupils, is obviously not comparable to one established for a chemistry test on a sample of twelfth-grade pupils; nor is a percentile norm for an English test based on a relatively unselected sample of ninth-grade pupils comparable to one established for a Latin test on a different and much more highly selected sample of pupils from the same grade. For these reasons, trying to interpret the scores of the pupil on a number of different subject tests has been something like trying to interpret a number of measures of the same physical object when its various dimensions are expressed in different units of measurement, and when little or nothing is known of the relationships existing among the various units employed.

In consideration of the lack of comparability and continuity in the measures involved, it follows obviously that no series of successive descriptions of individual achievement of the type obtained from subject examinations will provide any very meaningful indication of an individual's relative growth in the various aspects of his total development. The measurement of growth implies repeated measurement of the same trait, or repeated and periodical administration of the same or comparable tests of the same trait—something for which very little provision has been made (except in the elementary school) in educational achievement testing in the past. Scores on successive examinations in arithmetic, algebra, plane geometry, and trigonometry, for example, do not indicate how a pupil has improved in

his ability to do quantitative thinking in general. Nor do scores earned successively on a geography, a world history, a United States history, and an economics test (no matter how "comparable" the scores in a technical sense) indicate how much he has progressed in his thinking about contemporary social problems. Categories of subject matter are clearly not appropriate terms in which to describe educational growth and development. Single score measures of total achievement in individual school subjects cannot possibly measure well the progressive attainment of general educational objectives. Each of these scores is a measure of composite or average achievement with reference to a number of different objectives of instruction, and each is a different composite from every other. Satisfactory descriptions of total educational development can be secured only through tests each of which is designed to measure attainment of one general objective only, or one homogeneous group of objectives, each of which is constructed independently of the manner in which responsibility for this objective or set of objectives is distributed among the various school subjects, and each of which is administered periodically to all pupils, regardless of their current course registrations.

The Need for Tests of Hitherto Unmeasured Educational Objectives

If the descriptions of educational development of individual students provided by tests are to be truly comprehensive, tests and measuring devices must be developed for many more educational objectives than are now being measured at all. In general, satisfactory tests have thus far been developed only for objectives concerned with the student's *intellectual* development, or with his purely *rational* behavior. Objectives concerned with his nonrational behavior, or with his emotional behavior, or objectives concerned with such things as artistic abilities, artistic and aesthetic values and tastes, moral values, attitudes toward social institutions and practices, habits relating to personal hygiene and physical fitness, managerial or executive ability, etc., have been seriously neglected in educational measurement.

One of the reasons for the neglect in measurement of many of these objectives is that they have likewise been neglected in instruction, although either type of neglect may be regarded as a cause as well as a result of the other. A more potent reason is that attainment of these objectives is so difficult to measure, or that so little is known about how to measure them, just as so little is known about how to teach them effectively. (In general, the functional validity of tests will never far exceed the functional validity of instruction concerned with the same objectives, nor will the validity of instruction far exceed that of the tests.) Some of the reasons why these

objectives are so difficult to measure, as well as the general nature of the procedures that may be employed in their measurement, are suggested in the latter part of this chapter. Whatever the reasons, the fact and seriousness of this neglect are unquestioned. The present need for new tests of hitherto unmeasured objectives far exceeds the need for further refinements and improvements in existing tests.

Course Examinations in the Content Subjects

There is a real need in instruction (see pages 124–26) for examinations based on the immediate objectives and upon the peculiar organization and content of specific courses of instruction. While such examinations are severely limited in their usefulness so far as evaluation and guidance are concerned, they do serve effectively such purposes as the maintenance of teaching and learning standards (in assigning grades and determining promotion and failure), the motivation of students, and the evaluation of teaching effectiveness in relation to immediate objectives per se.

A basic premise in the preceding discussions is that the major uses of educational tests are in evaluation and guidance, and that, therefore, the principal concern of individuals and agencies constructing tests for wide-scale use should be with types of tests best suited to these major uses. This idea might be carried still further. The proposition is at least debatable that the construction of course examinations for the so-called content subjects should *not* be the job of central test construction agencies, that the wide-scale use of any specific examination of this type should not be encouraged, and that such examinations should not be employed in wide-scale testing programs. The corollary of this proposition is that the construction of specific course examinations in the content subjects should be primarily, if not solely, the responsibility of the individual classroom teacher and of the local school system. This proposition is undoubtedly as yet considerably premature—teachers are not yet sufficiently well trained in test construction to assume this responsibility, but a brief consideration of the reasons for the proposition should nevertheless be worth while here.

One argument for local construction of course examinations in most content subjects is that considerable variation from school to school in the content of such courses is not only permissible, but often highly desirable. This point was considered at length on pages 128–29 preceding, but there is still more to be said concerning it. In the social studies in particular, and only to a lesser extent in the natural sciences and the humanities, specific adaptation of the content and of the immediate objectives of instruction to local needs, interests, conditions, and facilities is clearly desirable. Some

variation in content should also be encouraged to take advantage of differences in the training, interests, experiences, and abilities of individual teachers. High school teachers of literature in a given state, for example, should not be required to teach exactly the same literary selections, but should each be encouraged to use the particular selections that they individually best like and best understand, and through which they can best impart to their pupils an appreciation of, and a liking for, good literature. In general, in all content subjects teachers should be permitted and encouraged, within limits, to use that content in which they individually are most interested, or with which they have had most personal experience and which they best understand.

Closely related to the preceding argument is the desirability of a high degree of *teacher participation* in the process of determining immediate course objectives and of selecting instructional content. If teachers should be encouraged to interpret general educational aims in terms of their own immediate objectives, and to experiment with different ways of attaining these objectives, clearly they should not be restricted by externally constructed content examinations imposed upon them by their administrative superiors. It should be noted, also, that the construction and improvement of their own course examinations constitute an unparalleled occasion for teachers to clarify their own thinking concerning the desired outcomes of instruction.

The direct measurement of ultimate objectives and the minimization of immediate objectives are, of course, just as desirable in locally constructed course examinations as in examinations intended for wide-scale use in evaluation and guidance. It is significant that the local test constructor enjoys certain important advantages in this respect over the constructor of tests intended for wide-scale use. This advantage may best be made clear through a specific illustration. Consider the problem of measuring the ability to comprehend literary materials. The constructor of a test intended for wide-scale use cannot assume that all of the examinees will be familiar with any particular literary selection. He must, therefore, limit his test questions to literary selections that are reproduced in the test itself, and for practical reasons he is obviously limited to a small number of very short selections. He is unable to raise questions calling for comprehension of very large units of content, or for comprehension, criticism, and evaluation of complete literary works. He is likewise unable to discover what the examinees best retain from their recent relatively free reading of complete works. The local test constructor, however, knowing that all of his examinees have recently shared certain reading and classroom experiences, may

take for granted temporary recall of the salient features of these experiences, and is free to raise questions of the type earlier suggested. Similar advantages in test construction are enjoyed by teachers of the social studies and natural sciences, who know that all of their examinees have had in common certain specific experiences—experiences far too complex to be reproduced, simulated, or described in the test situation—and can base their questions on these shared experiences.

The suggestion that a locally constructed course examination may be *based* upon the peculiar organization and content of the particular course involved, however, does not imply that the memorization of such content for its own sake is a legitimate end of instruction, or that such examinations should hold students responsible for that content for its own sake. The test questions may be "based upon" that content in the sense that they take for granted temporary recall of specific experiences (content) on the part of all examinees, but the questions should reveal differences among the students, not in the extent to which they have (temporarily) memorized the content, but in the extent to which they have derived the desired *lasting* outcomes from it. Since the recall is to a large extent only temporary, such examinations may rapidly deteriorate in validity after the students have completed the course, but when used *during* the course of instruction, such examinations may be definitely superior to any constructed for wide-scale use (see pages 128–29).

To take advantage of the aforementioned opportunities, the typical teacher must be very much better trained in the art and technique of test construction than he is at present. Until teachers are better trained in this respect, a considerable amount of centralized construction of course examinations for the content subjects seems desirable. In the long run, however, leaders in educational measurement should be more concerned with improving teacher competence in this respect than with continued production of course examinations for wide-scale use.

Summary

The principal points made in the preceding discussion may be briefly summarized as follows:

1. The contribution of educational measurement to education generally depends as much or more upon *what* test constructors elect to measure as upon how well they measure whatever they do measure.

2. Test constructors have, in the past, been too exclusively concerned with measuring the attainment of the traditional or prevailing immediate objectives of instruction in special school subjects. They have exhibited a relatively uncritical attitude toward these objectives, or have too often

accepted them without question as an adequate and authoritative basis for test construction.

3. Tests intended for wide-scale use, when based upon the most prevalent of the immediate objectives of current instruction, not only fail to measure well many of the significant outcomes of instruction, or to provide an adequate basis for educational guidance and the evaluation of instruction, but tend in themselves to perpetuate doubtful instructional and curriculum practices.

4. For purposes of individualization of instruction, guidance, and curriculum evaluation, much greater emphasis in test construction should be placed upon relatively direct measurement of the ultimate objectives of the entire educational program. For many of these objectives, no tests of any kind or quality are now available. Greater effort must, therefore, be made to provide more *comprehensive* descriptions of the students' general or total educational development at various levels, through *comparable* tests of such objectives. Such tests or test batteries should be planned and constructed quite independently of the present content and organization of school instruction, but in line with trends which are or should be developing in content and organization of instruction.

5. In practice, greater emphasis must be placed upon the periodic administration to all students, without regard to present course registrations, of comprehensive batteries of tests of the type just described. It is only in this way that comprehensive descriptions of the educational growth of individuals may be obtained.

6. Test constructors generally must assume much more responsibility than heretofore for the definition and clarification of ultimate educational objectives. Test constructors must exploit much more fully the potential values of tests and test construction techniques in the identification and clarification of objectives. (See chapter 2.)

7. The responsibility for the construction of course examinations in the so-called content subjects should be increasingly taken over by classroom teachers and local systems, and teachers in general should be better trained to assume this responsibility.

Basic Approaches in Educational Measurement

Direct vs. Indirect Measurement

Considerable emphasis has been placed in the preceding discussion upon the importance of measuring as directly as possible the ultimate objectives of instruction. The remainder of this chapter will consider the extent to which educational achievement can be directly measured, and will discuss

certain aspects of what might be termed the major strategy of achievement test construction, as contrasted with the more detailed and specific problems involved in writing the individual test exercises.

An educational achievement test may be described as a device or procedure for assigning numerals (measures) to the individuals in a given group indicative of the various degrees to which an educational objective or set of objectives has been realized by those individuals. Whether or not an educational objective has been realized in any individual can be ascertained only through his overt behavior. Indeed, in the last analysis, any educational objective is, in general terms, to condition or predispose the individual so that he will behave in a certain way in a certain situation. Accordingly, a test of any objective may be regarded as consisting in part of a situation or series of situations designed to elicit the desired behavior, or some other behavior which is presumably related to and will, therefore, predict the desired behavior, and in part of a procedure for assigning numerals to the properties of the behavior thus elicited, or to the product of that behavior.

The only perfectly valid measure of the attainment of an educational objective would be one based on direct observation of the natural behavior of the individuals involved, or of the products of that behavior. One of the objectives of high school instruction in the social studies, for example, might be "to so predispose the student that as an adult he will, at every opportunity, exercise the right to vote at elections of important public officials or on important public issues." The ultimate and conclusive test of the effectiveness of instruction in relation to this objective would require a tabulation of the number of times the individual had an opportunity to vote at such elections, and of the number of times he took advantage of this opportunity during his lifetime. This, and only this, would constitute a wholly direct, or perfectly valid, measure of the attainment of this objective. Numerous examples of what is meant by perfectly valid measurement, which are in every instance examples of direct measurement, are given in chapter 16 on "Validity," and, to avoid duplication, no further illustrations will be given here.

Direct measurement, then, is that based on a sample from the natural, or criterion, behavior series for each individual involved. Indirect measurement, on the other hand, may be defined as that based on behavior which is not a part of, but which is presumably related to, the criterion series. For example, one might attempt to measure indirectly the objective referred to in the preceding paragraph by noting the proportion of times that each high school student takes advantage of his opportunities to vote in school elections, on the assumption that those who most frequently exercise the

right to vote in these situations will, in general, be those who will most frequently do so later as adults.

It is only rarely possible, and even then not always practicable, to secure direct measures of the attainment of an educational objective for students yet in school. The reasons for this are fairly obvious, but a brief review of these reasons should nevertheless be worth while, particularly since they may contain some suggestions for the improvement of indirect measurement procedures.

In the first place, direct measurement of educational achievement is often impossible or impracticable because of the ultimate character of the objective involved, or because of the delayed appearance of the desired behavior. One illustration of this has already been given. While direct measurement of the "disposition to vote" objective is theoretically possible, obviously school teachers and counselors cannot wait for measures thus obtained, but must attempt by whatever indirect measures are available to predict *now* what each individual's criterion behavior may eventually be like.

A second reason why direct measure is often impracticable is that so often the natural behavior series is inaccessible to the examiner, or cannot readily be observed by him (for reasons other than the delayed appearance of that behavior). For example, one objective of instruction in high school science might be "to enable the student to make minor repairs of household mechanical and electrical equipment and installations." In this case, a part, at least, of the natural behavior series is accessible to the examiner, at least so far as time is concerned, although the complete series extends through the entire lifetime of the individual. Obviously, however, it is not at all convenient or practical for the examiner to make firsthand observations of the criterion behavior or to inspect its product.

A third reason why direct measurement of educational achievement is usually impracticable is the infrequency of current occasions for the specified behavior. To take an extreme example, the occasions on which a student in a life-saving class in swimming has to make use of what he has learned are altogether too few to base any useful measure of course achievement upon them. In this case, of course, a very close and satisfactory approach to direct measurement can be secured in simulated situations, but, unfortunately, this is not generally the case in the measurement of other types of achievement.

A fourth obstacle to the direct measurement of educational achievement is the lack of comparability in accessible behavior samples for different individuals. One of the objectives of elementary school language instruction, for example, is to develop in the pupils the habit of spelling correctly the words which they will be using in their own writing. This objective implies

not only that the pupils should spell correctly the words which they are now using, but also that as they continue to write more complex and varied materials they will spell correctly the new words needed. Here a sample of the products of the pupil's natural behavior is certainly immediately accessible. It would be quite feasible, for example, to collect practically everything that the pupils write during a given time period, say during one school semester, and to make an accurate count of the spelling errors committed by each pupil in his own writing. In this case, however, it would be extremely difficult to derive from these counts anything that might be regarded as comparable measures of spelling ability as defined by the ultimate objective. One pupil might be a very poor speller, yet he may now be attempting only a very limited amount of writing about simple subjects, using a very simple and restricted vocabulary, and may consciously or unconsciously be avoiding the use of words the spelling of which he is uncertain. Another pupil, who may be a very superior speller, may attempt to write at length about subjects demanding a very extensive and difficult vocabulary, and hence may create many more opportunities to misspell words than the first pupil. The number of spelling errors, even though expressed as a proportion of the number of running words or of the number of different words used, would hardly reveal the true differences in spelling ability in these pupils.

A fifth reason why direct measurement of educational achievement is seldom attempted is that it is often so costly in time and effort, or so inefficient, and hence may not be practicable even though it is otherwise possible. The foregoing example provides an excellent illustration of the inefficiency of direct measurement. To secure a count of spelling errors in the pupils' own writing, it would be necessary, in order to attain a minimum reliability of measurement, to examine thousands of words of writing for each pupil. Yet among the thousands of words examined, only an extremely small proportion would constitute spelling problems at all, and the great majority of the words inspected would contribute nothing at all to the aim of discriminating between good and poor spellers. Furthermore, the factor of illegibility of the pupils' writing would often be confused with that of actual misspelling and would serve, furthermore, to increase the time and effort required to make an accurate error count. The use of a list dictation test, in which the pupils write only words that are known to discriminate sharply between good and poor spellers, would obviously result in much more efficient and economical measurement. It is a fairly safe generalization that most "natural behavior series" consist in very large part of elements that are of zero or near-zero difficulty, or that are otherwise nondiscriminating for measurement purposes, and that relevant elements in the criterion

series are usually associated with many irrelevant, distracting, or misleading elements that tend seriously to lower the validity of the observations made for purposes of measurement.

A sixth obstacle to direct measurement of educational achievement, closely related to that just considered, is the relative complexity of most criterion behavior series, and the difficulty of analyzing out, or of isolating for observation and measurement, those elements of the total complex that are relevant to a given measurement purpose. In the natural behavior situations demanding arithmetic reasoning on the part of a ninth-grade pupil, for example, the reasons for his failure to arrive at a correct solution to a problem may often have little to do with his arithmetic ability as such, but may be the result of distracting influences, of emotional factors, and so forth, which may be extremely difficult to identify and to separate from the factors involved in arithmetic reasoning per se. How much more difficult it would be to identify in the natural behavior of the adult those elements of "good citizenship" which are attributable to a particular course of school instruction is too obvious to warrant discussion.

THE FOUR BASIC TEST TYPES

The implications of the preceding discussion are fairly clear. Instead of observing the examinee's behavior in a sample of the situations which *present themselves* to the examinee in the natural course of events, the examiner must in most cases[2] present *to* the examinee a number of situations especially selected or designed to elicit such behavior. It should be noted that if it were possible to observe the entire criterion series for any two individuals, those series would almost certainly differ markedly for those individuals. Accordingly, the selected test situations, which are the same for all examinees, may not strictly be regarded as a random or representative sample from the criterion series for any examinee. Furthermore, for the sake of expediency, efficiency, and comparability, it is necessary to eliminate from, or to control in, the test situations many irrelevant factors or nondiscriminating elements which would operate or be present in the criterion situations. Because of these restrictions and controls, the test situations become relatively artificial in character, and hence the validity of the test may never be taken entirely for granted. The problem for the test constructor, then, is to devise a test series that will be *as much like,* or as *closely related to,* the criterion series for all examinees as considerations of expediency, efficiency, and comparability will permit.

[2] Limited use is made, in educational measurement, of observation check lists, questionnaires regarding previous behavior, anecdotal records, and records of specific activities such as reading records, and the like.

Accordingly, the most basic classification of tests into types is that which depends upon the nature of the relationship of the behavior comprising the test series to that constituting the criterion series. Four such basic types of tests may be identified. It may be noted, first, that all educational objectives are ultimately concerned with behavior, that is, with what the examinee can or will *do* in specified situations. Accordingly, the test constructor may (1) give the examinee special occasion to do some of the things that are specified by the objective (without waiting to observe those things in the natural course of events), and assign measures on the basis of the frequency or adequacy with which he does those things; (2) give the examinee occasion to do things similar to some of those specified by the objective, and assign measures on the basis of the assumed relationship between the behavior elicted by the test and that constituting the criterion series; (3) describe the situation in which the examinee would have occasion to do what the objective specifies, and then ask him to tell what he would do in this situation or how he would do it, and (4) discover whether or not the examinee knows the facts, rules, principles, etc., that are presumably essential or conducive to the desired behavior.

There are relatively few instances of actual tests that are purely and solely of one of these types only. Often the same test may exhibit some of the characteristics of each of these four types. There is certainly no implication that tests should be purely of one type or the other, nor is it suggested that any effort should be made to classify existing tests on this basis. The distinctions are here made primarily for the purposes of this discussion, in an attempt to develop a clearer understanding of certain basic problems in test construction. For these purposes it should be worth while to illustrate and consider briefly some of the more important characteristics of each of these types.

The "Identical Elements" Type of Test

The first of these four basic types may, for want of a better name, be called the "identical elements" type. The most important characteristics of this type are that certain behavior situations are presented to the examinee for the special purpose of measurement, and that the elements of behavior elicited by the test situations are practically identical with certain of the elements comprising the criterion series for the individuals involved. The test series, however, does not contain all of the elements comprising the criterion series; rather, only the more discriminating, or the more readily reproducible, or the more crucial, or the more readily measurable, or the more relevant of the elements of the criterion series are selected for the

test, and these elements may, in important respects, be quite differently distributed than in the natural or criterion situations.

A good example of a test that meets this description is the type of stenography test that is frequently used in business schools. The test content may consist of a specially prepared business letter which is in many respects representative of the letters that the student may later have occasion to type as an office stenographer and typist. This letter is dictated to the class of students as a group, and the students take the letter in shorthand and then prepare a typed copy of it from their own stenographic notes. The letter is scored by assigning predetermined weights to such things as the number of typing errors made in predetermined error situations, the arrangement of the letter on the page, the neatness and uniformity of the carbon copies, and the time required to type the letter. The letter may be far from typical of actual business letters in that, for the sake of efficiency of measurement, it may be much more heavily loaded with opportunities for certain types of error than a typical letter would be, and many elements frequently found in business letters but nondiscriminating for measurement purposes may be entirely omitted from the test letter; yet the elements that are present in the test letter are, for all practical purposes, identical with those found in letters typed in the "natural behavior" situations. The test situation is, of course, a somewhat artificial one; many of the attendant circumstances may be quite different from those associated with the natural behavior series, but for the purposes at hand these differences or artificialties may be of no practical consequence.

Most so-called "work sample" or "performance" tests are predominantly of the identical elements type. For example, the performance tests sometimes given to applicants for automobile driver's license, or those given in shop courses in industrial and domestic arts, clearly fall in this category. (See pages 456–63 for further illustrations.) While these tests often contain irrelevant elements which may bias the measures obtained, nearly all of the things the examinees are required to do are practically identical with those that they would be expected to do in the natural situation.

A particularly important and widely used test that is primarily of this type is the test of reading comprehension. Such tests typically consist of a collection of reading passages, much like those which the student would later have occasion to read and interpret in the real life situation, and which are accompanied by a series of questions requiring the student to derive meanings and draw inferences like those that he might have occasion to derive in his own free reading. As actually constructed, such tests are often of low validity due to poor selection of reading passages and to failure to

devise questions that will require the students to exercise some of the more important components of general reading ability. In a well-constructed reading test, however, most of the things the student is required to do are essentially the same as those he does in the natural situation. In other words, the best reading test in itself constitutes the best available definition of what is meant by reading ability.

Incidentally, it is in tests of this general type, particularly in tests of ability to *interpret,* to *evaluate,* and to *think critically* about complex reading materials, that the art of educational achievement test construction has perhaps reached its highest level of development to date.

The "Related Behavior" Type of Test

The second basic type of test is that in which the elements of the test series are not a part of the criterion series, but are presumably substantially related to the criterion behavior for the population involved. The elements of the test series may or may not be similar to that in the criterion series, the essential condition being that measures based on the two series be highly correlated. In educational achievement testing, however, for reasons later to be considered, the aim is usually to secure and insure the desired correlation by making the test series as much like the criterion series as possible.

In actual practice, it is impossible to draw any sharp line of distinction between the identical elements and the related behavior types of tests. Actually, some elements of the test series are usually identical with those in the criterion series, but many of the elements in the test series are not present at all in the criterion series, the proportion of such elements varying considerably from test to test.

One example of a test of this type has already been suggested—that in which the adult's voting behavior is predicted by his voting behavior in the student government situation. A test in which the general trait of "honesty" is measured by determining how frequently students cheat in a specially devised examination is another example of this type. Most so-called "simulated behavior" tests belong to this category, such as tests of the type used by the military services in personnel selection in the last war in which, for example, the examinee might sit in a mock cockpit behind a mock machine gun and "fire" at images of pursuit planes on a moving picture screen.

The varying proportion of identical elements that may be found in tests of this type is readily illustrated in the measurement of spelling ability. As has already been noted, measures of spelling ability might be derived from error counts based upon the free writing of the students. Such a test would be almost purely of the identical elements type. To control the sampling of spelling opportunities from pupil to pupil, however, the device

might be employed of dictating the same sentences to all pupils, and counting the spelling errors in these sentences. In this case, the score of any pupil may depend, in addition to his spelling ability, upon such things as how clearly the examiner dictated the sentences, how accurate is the hearing and how close the attention of the examinee, how rapidly the sentences were dictated, how many clues to the correct syllabification and spelling are provided by the articulation or pronunciation of the words dictated, how adequately the context distinguishes some of the words from their homonyms, etc. Greater efficiency may be obtained by dictating a list containing only words known to be difficult or discriminating, but only at the cost of increasing the artificiality of the test situation. It is quite conceivable, for instance, that in a test situation of this type, where the pupil's attention is sharply focused on the problem of spelling particular words, and where he is helped by hearing the words pronounced, his spelling of some words will differ from that which he would habitually and unconsciously employ in his own free writing. In the list dictation test also the manner in which the words are dictated may again introduce irrelevant variations in the test scores.

Still greater efficiency may be secured, and some of these irrelevant factors may be controlled, by providing the pupils with printed sentences or lists containing misspelled words, with instructions to write the correct spelling above each misspelled word. From this, it is but a short step to the type of test in which the examinee is presented with a number of groups of printed words, and asked merely to identify the misspelled word in each group, without having to provide the correct spelling. In tests of this type, the relative frequency with which the examinee detects and corrects, or only detects, spelling errors in printed sentences or lists prepared by others, and the relative frequency with which he would habitually and automatically spell the same words correctly in his own writing, may be far from perfectly correlated.

The foregoing example illustrates the general truth that increased efficiency and economy in test administration and scoring and increased control of irrelevant factors are usually secured by introducing greater and greater differences between the behavior actually tested and that constituting the criterion series. In order to control irrelevant variations of some types, other irrelevant factors must be deliberately introduced. Greater comparability and greater reliability per unit of time are thus achieved, but frequently at a definite sacrifice in intrinsic validity.

The "Verbalized Behavior" Type of Test

The third basic test type is that in which selected behavior situations from the criterion series are described to the examinee, and in which he

tells how he would (or how one ought to) behave in those situations. The description or presentation may be either verbal (oral or written) or visual (pictures, charts, diagrams, or movies). The examinee may be required to tell what he would do, or how he would do it, in his own words (written essay or oral), or he may select the one of a number of suggested descriptions of possible behavior which he considers correct or best (multiple-choice test). This type of test is really a variation of the related behavior type, but is sufficiently important to deserve separate consideration.

An example of this type of test would be that in which teacher candidates are presented with a series of detailed descriptions of classroom situations in which problems of classroom management have arisen, in which each situation described is accompanied by descriptions of various actions that the teacher might take in handling the situation, and in which the examinees are to select the action that in their judgment is best for each situation.

Another test of this type would be that in which medical students are presented with a series of descriptions, possibly supplemented by photographs, X-rays, etc., describing the symptoms of a number of patients, each description being accompanied by several suggested diagnoses and several suggested therapies, the task for the examinees being to select the correct diagnosis and correct therapy for each case.

Still another example of this type would be a test in which students in a high school course in Family and Marriage are presented with descriptions of a number of domestic problem situations, each of which is accompanied by a description of several solutions to the problem which might be adopted by the persons involved, the examinees being required to select the description that they consider represents the best solution in each case.

The foregoing examples have all been of tests of the objective type; the manner in which the same idea can be utilized in tests of the essay type is obvious.

For a great many educational objectives, the criterion behavior series is of such a character that it is utterly impractical to attempt to reproduce or to simulate in the school examination the *overt* behavior with which the objective is ultimately concerned. In many such situations, the verbalized behavior type of test may serve as a very acceptable and practicable substitute for the overt behavior type, and might do so even though the latter type could be employed. In the social studies, particularly, this is a very promising test type, and one whose potentialities test constructors have only begun to exploit.

The Test of the Student's Knowledge

The fourth basic type of test is that which seeks to discover how much the student knows about a particular topic or subject. It has been perhaps the most widely used of the various test types and is too well known to require specific illustration here. The majority of school examinations or standardized tests in most of the so-called content subjects have been almost exclusively of this type.

The limitations and possibilities of this type of test have already been considered at length in the earlier part of this chapter and ways of improving tests of the knowledge type generally will be extensively considered later in chapter 7, "Writing the Test Item." It will be sufficient for the present purposes, then, only to remind the reader of a few of the most important characteristics of tests of this type.

In the first place, in evaluating tests of this type, it is especially important to distinguish clearly between tests of "knowledge" in the sense of understanding or of acquiring *meanings*, and tests of knowledge in the sense of *verbalizations* that have been learned by rote and are practically devoid of meaning and functional value to the learner. The acquisition of meaning is an active process, involving the drawing of inferences, the relating of items, the translation of statements into one's own words, the formulation of generalizations, the finding of illustrations and applications, and so forth. At the one extreme, a test of the student's knowledge may require the student to do these things, and may be primarily concerned with relatively broad concepts, basic principles, fundamental generalizations and relationships. At the other extreme, a test of the student's knowledge may be concerned only with poorly selected, detailed, descriptive facts and may require nothing more from the examinee than the recall or recognition of verbal stereotypes which have been memorized but not understood. Unfortunately, a large proportion of educational achievement tests have approached closer to the latter extreme than to the former.

There is, of course, very little justification for tests of the latter type in educational achievement testing. The justification for tests of the student's knowledge in the more desirable sense rests upon the contention that extensive knowledge is essential or conducive to the overt behavior with which the ultimate educational objective is concerned. In other words, the final justification for this type of test is that there is high correlation between how much or what an individual knows and how he will behave in certain situations. (Thus, the knowledge type of test may also be regarded as a variation of the related behavior type.) This justification undoubtedly has some validity, and tests of the student's knowledge in the more desirable sense will undoubtedly always play an important role in

educational achievement testing. However, while extensive and accurate knowledge may be an essential condition to effective thinking and to desirable overt behavior in certain situations, it is by no means a sufficient condition. The correlation between test behavior and criterion behavior for this type of test is therefore by no means perfect, and is particularly likely to be quite low where acquisition of knowledge for its own sake has become an end of instruction, as it so frequently has in practice. Tests which measure how well the student can make use of what he knows (and which thus indirectly measure also how much he knows) are therefore much to be preferred to tests of knowledge alone.

The Fundamental Goal in Achievement Test Construction

It has been shown that, for various reasons, direct measurement of ultimate attainment of educational objectives is, in most instances, impossible or impracticable in the school situation. At best, the educational test constructor can elicit in the test situation only a limited sample of the behavior elements constituting the complete criterion series. Usually, he must be content to substitute for many elements of the criterion series definitely dissimilar (but presumably related) elements of behavior that are immediately and readily accessible to him. Nevertheless, *it should always be the fundamental goal of the achievement test constructor to make the elements of his test series as nearly equivalent to, or as much like, the elements of the criterion series as considerations of efficiency, comparability, economy, and expediency will permit.* The more nearly the test itself completely defines the ultimate educational objective involved, the more satisfactorily the test will serve its many purposes. The aim of the test constructor is thus always to make his test as much of the identical elements type as he possibly can, and to resort to the use of the other types only when no other procedure is at all practicable.

It may sometimes be possible, with a test series that differs considerably in character from the criterion series, to demonstrate quite conclusively that a high relationship exists between measures based on the two series for the population for which the test is intended. It is very important to observe, however, that this in itself does not constitute adequate justification for the widespread and continued use of the particular test involved. This is because the widespread and continued use of a test of the character just described will, in itself, tend to reduce the correlation between the test series and the criterion series for the population involved. Because of the nature and potency of the rewards and penalties associated in actual practice with high and low achievement test scores of students, the be-

havior measured by a widely used test tends in itself to become the real objective of instruction, to the neglect of the (different) behavior with which the ultimate objective is concerned. The students attain greater proficiency in doing what the test requires, without improving themselves in the criterion behavior, and the correlation between the two is lowered. Thus, the test may not only exercise an undesirable effect upon instruction and learning, but the validity of the test itself deteriorates more and more as the test is more widely used. This may occur not only with tests of the related behavior type, but with tests of the identical elements type as well. In the latter type of test, not all of the elements of the criterion situations are included in the test series, and those not included tend to be neglected in instruction, or learning, or both. The effect, of course, is most serious with tests of the related behavior type. Thus, tests concerned only with verbalized behavior are likely to encourage undue emphasis upon verbalization in learning, or tests concerned only with the extent of the student's knowledge are likely to cause neglect in instruction of the functional values of that knowledge to the learner.

The practice of cramming for educational tests, although perhaps not a very serious problem generally, is of interest here because it serves to bring into sharper focus the problem just raised. Cramming for tests is really undesirable only to the extent that it fails to promote, or actually interferes with, the attainment of the ultimate objective with which the test is concerned. If the test encourages intensive but temporary memorization of ill-digested facts or meaningless verbalizations, for example, then, of course, the result is bad. If, on the other hand, the test encourages the student to do what he ought to do in any event, that is, if the test series is identical with the criterion series, then cramming is all to the good, and the more of it the better. Interestingly enough, the more nearly the test itself adequately defines the educational objective involved, the less does cramming for the test tend to be practiced at all. For example, there is never much of a problem of cramming for a good reading comprehension test. Students and teachers soon learn that the only way to secure high scores on such tests is really to improve in reading ability, something which they readily acknowledge that they do not know how to accomplish in the few hours preceding the examination.

The principal danger in achievement test construction, then, is that the behavior comprising the test series may not be sufficiently representative of the entire criterion series. As has already been noted, this is true of the identical elements type of test as well as of the others. In constructing tests of the identical elements type, the tendency is to limit the test series to the

elements of the criterion series that are most conveniently and most easily reproduced, or most easily and objectively observed and evaluated, and many of the more unmanageable but more important and crucial elements tend to be neglected in, or omitted from, the test. Similarly, in constructing tests of the related behavior type, the tendency is to simulate or to reproduce verbally only those elements of the criterion series which can be most easily simulated or reproduced.

Complex vs. Simple Tests and Test Exercises

There is one element that is common to a great many criterion series, which tends particularly thus to be neglected or ignored in the construction of achievement tests. Attention has previously been drawn to the great complexity that frequently characterizes the behavior comprising the criterion series for an ultimate educational objective. This complexity may in itself constitute the very essence of the criterion behavior, or may be regarded as the most important single aspect of, or element in, that behavior. It is rather common practice in achievement test construction to attempt to analyze a complex criterion behavior into relatively pure or simple elements or traits, to attempt to measure each of these elements separately or independently, and then to combine these measures into a single composite score for the purpose of the action judgment which must eventually be made. This tendency has possibly been encouraged by the emphasis placed upon analytical or factorial analysis procedures in aptitude and psychological testing. In certain respects, this influence may have been an unfortunate one. Eventually, test theory and technique may advance to the stage where a composite measure thus obtained will accurately describe or evaluate the complex behavior of the criterion series, but, for the present, it seems best to attempt to incorporate in the achievement test situation as much as possible of the same complexity that characterizes the criterion situation. This is particularly true in tests of critical thinking or tests of the ability to interpret and evaluate complex materials in the social studies, the natural sciences, literature, etc. In such tests the most important consideration is that the test questions require the examinee to do the *same* things, *however complex,* that he is required to do in the criterion situations, even though it may consequently be very difficult to classify these questions into clear-cut homogeneous categories, or to establish meaningful part scores for such categories. In building a reading comprehension test, for example, one can be too much concerned with the attempt to have certain items measure only the "ability to note details," others measure only the "ability to organize ideas," and still others only the "ability to infer meanings of words from context," to the extent that

he may fail to employ more valid questions which have occurred to him, but which he excludes because they are too complex in character to be readily classed in categories of the type suggested. As a result, the essence of the complex behavior may be analyzed out, because the sum of the parts may not be equal to the whole.

Closely related to the problem of reproducing in the test series the essential complexity of the criterion series is that of reproducing or simulating in the test series the full scope and variety of the criterion series, or of including in the test all of the relevant components of the natural behavior situation in their proper relation to one another. The tendency is greatly to oversimplify the test situation and to exclude from the examinee's consideration many specific elements that might prove crucial in determining his behavior in the corresponding natural situation. In tests of the verbalized behavior type, in particular, the tendency has been to make the descriptions provided altogether too short to accomplish their purpose well. In actual classroom management situations, for example, the action taken by the teacher would often depend upon a very intimate acquaintance with the personalities and past performances of all of the individual pupils in the group, and only a very skillful examiner could succeed, even in several pages of description, in presenting many of the more subtle factors that would influence the teacher's judgment in the real situation. Likewise, it would be extremely difficult to anticipate in the medical diagnosis test all of the characteristics of the patients which might be regarded by some examinees as symptoms of a particular disorder, and on which their diagnosis might hinge. Frequently, in the real situation, one of the major difficulties involved in arriving at a sound judgment is that of identifying the elements in the total situation that are truly relevant to the particular problem in hand. The ability to distinguish relevant from irrelevant factors is thus an important part of the total ability to be measured, and the situation described in the test must therefore contain many elements which are irrelevant to the solution of the problem presented to the examinee, but highly relevant to the purposes of the test.

Still another aspect of the need for more complex, as well as for simpler, tests and test exercises deserves specific mention. Great emphasis has been placed, in discussions of the curriculum and of methods of instruction, on the need for closer coordination or integration of instruction, and on the desirability of giving students more occasion to use together, in an integrated attack on complex problems, the many skills, knowledges, and abilities which they have acquired in different school subjects and at different times. Tests that may be used to evaluate the extent to which instruction has been effectively integrated, and that will place an effective premium

upon such integration, are as sorely needed as integration of the content and methods of instruction of itself. Obviously, only tests of a highly complex character can adequately fill this need.

In connection with this discussion of test complexity, it may be noted that for many practical purposes in education, too many test scores may be worse than too few. It would be quite possible to analyze a high school student's total educational development into, say, a hundred specific elements or aspects. Few teachers or counselors, however, would be capable of interpreting such a mass of test data, particularly in view of the probable lack of comparability in the many measures presented. In general, such data, however numerous, are used as the bases for a very small number of critical action judgments. Since any action judgment must in any event be based upon no more than a few composites, it might very often be better to use as those composites the scores on a few realistic and complex tests, rather than the weighted combinations of scores on a very much larger number of simpler and more homogeneous subtests which do not adequately define the whole of what the tests are intended to measure.

The foregoing is certainly not meant to imply that there is no place for diagnostic or analytical testing in educational measurement. On the contrary, the description of an individual provided by test scores should always be as analytical a description as can be advantageously used for the practical purposes in hand. A reading test intended for purposes of general educational guidance, for example, might best be of a complex character, while one intended to serve as a basis for diagnosis leading to remedial reading instruction with problem children may have to consist of quite a number of relatively homogeneous parts. Even here, however, teachers must be cautioned not to assume that the desired whole will necessarily result automatically from isolated attacks on the parts identified by the test.

The Limitations of Written Examinations

Reference has been made earlier to the fact that many of the more important objectives of the whole program of general education have thus far received little or none of the attention of measurement workers. This fact is closely related to the extremely heavy reliance that has been placed upon pencil-and-paper techniques of measurement in education. Written examinations, in general, are well adapted only to the measurement of the *intellectual* aspects of the student's educational development, or of his *rational* behavior. The verbalized behavior type of test, for instance, may often serve as a fairly satisfactory substitute for direct measurement of the

overt behavior specified by an ultimate objective, but only if that behavior is almost completely determined by logical considerations. For example, a well-constructed medical diagnosis test (see page 150) might work quite well in this regard, but even the most skillfully constructed test for a course in Family and Marriage (see page 150) might fail to predict well the actual behavior of the examinees in the corresponding natural situations. In the examination situation, the student may profess certain attitudes or beliefs, or indicate that he would act in a certain fashion, simply because he knows that certain responses are socially approved, and that only such responses will be given credit in scoring the test.

For many objectives of the type suggested, measuring or observational devices concerned with the student's overt behavior while he is yet in school, even though of a rough and opportunistic character, are perhaps much more worth while than any written types of tests. The actual behavior of the student in school elections, for example, however fragmentary or limited in sampling, may provide a better clue to his future behavior than anything he professes that he will do in a written test; books actually read by the pupil in his free time may constitute a truer indication of his literary tastes than his score on a literary appreciation test; anecdotal records may be more meaningful than scores on personality tests; etc. The problem of deriving comparable measures from such opportunistic observational data now seems to present almost insuperable difficulties, but perhaps no worse than have been resolved before through determined and persistent effort.

This volume—especially that part of it concerned with test construction—deals almost exclusively with written examinations. This is because it is only with such examinations that measurement workers in education have acquired any large body of experience and knowledge that can be handed on in meaningful form to the beginning student. Many of the principles of test construction, however, and most of the theory of measurement herein presented, are generally applicable to any and all types of tests and measuring devices. It is much to be hoped that, in the future, educational measurement workers will apply these principles and suggestions to the many areas now so sorely in need of attention.

Conclusion

Most of what has been said in this chapter represents expression of opinion only. Few if any of the statements made could be closely documented or conclusively substantiated by concrete experimental evidence. Some of these statements would be subscribed to by practically all leaders in this field; on others there might be sharp disagreement. The one thing

on which all certainly would agree is that the questions here considered represent the kinds of questions to which measurement workers generally should devote a much greater share of their attention. If measurement is to continue to play an increasingly important role in education, measurement workers must be much more than technicians. Unless their efforts are directed by a sound educational philosophy, unless they accept and welcome a greater share of responsibility for the selection and clarification of educational objectives, unless they show much more concern with what they measure as well as with how they measure it, much of their work will prove futile or ineffective.

The author's primary purpose in this chapter, therefore, has not been so much to suggest answers to the questions raised as simply to raise the questions, or to indicate the general direction of the thinking which should precede the selection of any specific test construction project. If the answers that have been suggested may at times have seemed dogmatic in character, this may itself contribute to the central purpose of stimulating more critical consideration of the questions raised.

6. Planning the Objective Test

By K. W. VAUGHN
Formerly with *Cooperative Test Service*

COLLABORATORS: Dorothy C. Adkins, *University of North Carolina;* Louise Witmer Cureton, *University of Tennessee;* Frederick B. Davis, *Hunter College;* Geraldine Spaulding, *Educational Records Bureau*

PLANNING IS AN ESSENTIAL ACTIVITY IN ALL STAGES OF A TEST construction project. Inattention to planning may result in a failure to meet production deadlines, or may necessitate the use of uneconomical procedures or of below-standard materials in order to meet those deadlines. It may mean that certain desirable procedures will prove unusable in the later stages of the project because inadequate foundation was laid for them. It may result in the wasteful preparation of more items than are needed in the finished test, or in a failure to prepare enough materials to survive the tryout. In general, it may lead to countless annoyances and delays due to a failure to coordinate properly the various phases of test production.

Test planning encompasses all of the many and varied operations that go into producing a test. Not only does it involve the preparation of an outline or table specifying the content or operations to be covered by the test, but it must also involve careful attention to item difficuly, to types of items, to directions to the examiner, to arrangements for tryout, to problems of test reproduction, to provision for expert review, to the provision of adequate equipment and facilities, to the procurement of personnel, and so forth. These are only a sample of the many operations in, and aspects of, test construction that demand planning. This chapter thus provides a brief orientation to a number of problems that are more thoroughly discussed in later chapters. The major concern of the chapter, however, is not with the details of how the various difficulties involved in test construction are to be met, but rather with the need for anticipating these difficulties before they arise, of coordinating and tying together the various operations involved, and of insuring smooth and efficient administration of the project as a whole.

Defining the Purpose of the Test

Most test construction projects are originally undertaken with only a general or somewhat indefinite conception of their purpose in the minds of the test constructors. Perhaps the most basic step, therefore, is to analyze and clarify these general objectives so that the purposes of the test can be stated in specific, concrete terms.

For some types of tests (called "predictor" tests in chapter 9), satisfactory criterion data may make possible the computation of meaningful validity coefficients. In such a case the general statement of the purpose of the test might be to predict a particular criterion, and the final check on the test would be its relationship to this criterion. For most achievement tests, however, meaningful validity coefficients cannot be obtained. The validity of a test of this kind can be estimated only by subjective judgments regarding the extent to which it measures what it is intended to measure.

For either type of test, there should be available a general statement of the test purpose. Such a general definition of objectives should specify what is to be measured, who is to be tested, and what uses are to be made of the test scores. The purpose of a French test, for example, might be stated as follows:

> French, as defined for the purposes of this test, consists of knowledge of French words, ability to translate French prose, knowledge of French grammar, and information about France and French culture. This test is to be administered to public school pupils throughout the United States who are completing two to six semesters of French and who are in grades nine, ten, eleven, and twelve. The resulting scores are to be used as one factor in assigning grades.

So general a statement, of course, serves merely as a starting point, and must be broken down into a much greater detail to provide a meaningful guide to the item writer.

In the case of predictor tests for which a criterion measure is available, use of a carefully prepared outline of test content is the best guarantee that all measurable relevant abilities will be covered. Before the items are actually correlated with the criterion, the judgment of the examiner must be applied in attempting to secure adequate coverage. The statistical analysis serves as a check on his judgment. For self-defining achievement tests, a fairly comprehensive outline of test content is perhaps even more essential.

Sometimes the statement of purpose for a test leads at once to the formulation of a test outline; at other times considerable labor is involved. Particularly for educational achievement tests of the self-defining type,

the detailed definition of test purpose may depend upon the analysis of behavior, of jobs, of textbooks, or of curriculums, and may entail time-consuming consultation with subject-matter specialists. In some cases the definition of purpose may be confined to a fairly informal sketch of topics to be covered, particularly when the purpose is very narrow, or when it is prepared for highly competent test constructors. The labor in preparing the test outline varies with the nature and purpose of the test and with who is to build it.

Preparing an Outline of Test Content

Two Aspects of Test Content

The term "test content" has been used rather broadly to cover both the subject matter of the test or indivdual test item and the type of ability that it is thought to require.

Sometimes these two aspects of test content can conveniently be treated together. For example, if "ability to compute the arithmetic mean from grouped data" is listed as a test topic, both the subject matter and type of behavior to be tested are at once clear.

On the other hand, often the specific content might well be separated from the type of behavior to be tested and attention be directed specifically to each aspect separately. Too often a test outline is confined to subject matter alone, and the type of behavior to be tested is left to the judgment of an inexperienced item writer. One of the later illustrations shows a convenient way of segregating both aspects of test content in a two-way table (page 162). Thus, a test outline can be used to show clearly not only what different areas of subject matter are to be covered but also the types of behavior to be elicited with respect to each area. As will be seen, it can also indicate the relative emphases of the various topics and of the several types of behavior. Moreover, it can reveal the relative emphases on different types of behavior for each subject-matter area, as well as the relative emphases on different topics for each type of behavior.

Ways of Determining Test Content

Analysis of objectives of measurement

Since the fundamental objectives of education are ultimately concerned with the modification of human behavior, it is advisable to analyse the objectives of instruction to determine what activities and skill should be appraised in an educational test. The practical difficulties that often prevent satisfactory measurement of the ultimate objectives of education are

described in chapter 5 and will not be repeated here. All objectives preferably should be included in the test outline. Explicit note should be made that any particular objective cannot be or is not being measured.

An illustration of the development of a relatively broad test outline is provided by the general plan for the 1948 form of the Premedical Science Achievement Test of the Association of American Medical Colleges (*5*). The items were distributed as follows by objectives of instruction and subject-matter fields:

	Percent of Items
1. Objectives of instruction	
a) Comprehension, interpretation of scientific materials	35
b) Application of concepts, principles, etc.	30
c) Recall of basic concepts	35
Total	100
2. Subject-matter field	
a) Biological sciences	40
b) Chemistry	40
c) Physics	20
Total	100

Combined in a two-way table, these specifications led to the breakdown shown in Table 5.

TABLE 5

NUMBER AND PERCENT* OF ITEMS IN EACH CATEGORY OF THE 1948 PREMEDICAL SCIENCE ACHIEVEMENT TEST†

Subject-Matter Field	Objectives of Instruction‡							
	Recall of Basic Concepts		Application of Concepts, Principles, etc.		Comprehension, Interpretation		Total	
	N	Percent	N	Percent	N	Percent	N	Percent
Biological sciences	18	*12*	15	*10*	26	*17*	59	*39*
Chemistry	21	*14*	25	*17*	14	*9*	60	*40*
Physics	4	*3*	14	*9*	13	*9*	31	*21*
Total	43	*29*	54	*36*	53	*35*	150	*100*

* Rounded to the nearest whole number.

† Reproduced from an unpublished report to the Association of American Medical Colleges by the Graduate Record Office, December 1947.

‡ Items classified in more than one category are assigned fractional weights in these categories.

It should be made clear that in practice a much more detailed outline of the contents within each cell of a table such as Table 5 is needed before test construction proceeds. If such an outline is not prepared by persons thoroughly acquainted with the subject-matter fields and with the purposes of the test, too great a share of the test planning is likely to be done by the item writers. Even if the item writers are equally competent in the

subject-matter fields, detailed breakdown of the content to be covered at the stage of test planning almost always is worth while in insuring adequate coverage and in avoiding uneconomical construction of too many or too few items on a particular topic or requiring a particular type of behavior.

In practice it may not be feasible to attempt to adhere very rigorously to the weights indicated in the cells of a table such as that presented in Table 5. Frequently the same item may be classed in two or more cells, or ideas for good items may be more plentiful for some cells than for others, or more items may survive the tryout in some categories than others. Furthermore, as will be made clearer later, the weight actually given to a category is by no means a function only of the number of items assigned to it, and the number of items indicated in the table is only a very rough estimate of the number really needed to secure the desired weight. Accordingly, the weights indicated in the original table may be altered somewhat during the course of the test construction, as sound reasons for doing so are encountered. Tests intended to measure achievement in particular courses of instruction must, to some extent at least, be based upon what the pupils were actually taught, rather than upon what someone may think should be taught. However, the limitations that this procedure imposes on the test so far as its uses in evaluation are concerned must be clearly recognized (see pages 123–34).

A second illustration, confined entirely to the analysis of behavior objectives, is provided by the following list prepared in the course of planning the college chemistry test of the United States Armed Forces Institute. There is no standard introductory course in chemistry offered by most colleges and universities, but there is a fairly universal agreement on certain topics that should be taught in more or less the same sequence. Furthermore, the objectives of instruction in chemistry are reasonably well agreed upon. The outline for the USAFI examination in college chemistry, therefore, was designed to link the basic objectives of instruction to the subject matter generally taught in most colleges and universities (*2*).

Behavior Objectives

I. Ability to recall important information
   A. Knowledge of important facts
   B. Knowledge of definitions of important terms
   C. Acquaintance with important concepts
   D. Verbal understanding of theories and principles
   E. General knowledge of the physical properties and chemical behavior of the more important elements and their compounds

II. Ability to apply principles in making simple predictions
  A. Functional understanding of the principles and theories of chemistry and their interrelationships
  B. Application of a definition
  C. Application of a principle in situations similar to those encountered in a typical course
  D. Application of principles in new situations taken from everyday life
  E. Interpretation of a set of data and drawing conclusions from them

III. Ability to apply principles quantitatively by carrying out calculations
  A. Quanitative significance of chemical symbolism
  B. Balancing of chemical equations

IV. Ability to use the scientific method
  A. Distinction between observed phenomena and their theoretical explanation
  B. Explanation of phenomena in terms of theory
  C. Presentation of the experimental evidence for a theory
  D. Identification of the assumptions underlying a given conclusion
  E. Identification of the factors that must be controlled in an experiment
  F. Identification of statements that are true merely by definition

Such an analysis as the foregoing of behavior objectives should prove very useful in the preliminary stages of planning a test. Before actual construction of items begins, however, it should be accompanied by an equally comprehensive analysis of the content to be sampled.

A third example will be drawn from an operational rather than a content field. As a preliminary step in constructing a reading test, Frederick B. Davis (*3*) surveyed the literature to identify the comprehension skills deemed by authorities to be most essential in reading. A list of several hundred specific skills was made, and these were then classified in an attempt to group together skills that are closely interrelated and that have low correlations with skills in other groups. From this approach there emerged the nine groups of skills shown below.

Davis' Classification of Reading Skills

1. Knowledge of word meanings
2. Ability to select appropriate meaning for a word or phrase in the light of its particular contextual setting
3. Ability to follow the organization of a passage and to identify antecedents and references in it
4. Ability to select the main thought of a passage
5. Ability to answer questions that are specifically answered in a passage
6. Ability to answer questions that are answered in a passage but not in the words in which the question is asked
7. Ability to draw inferences from a passage about its contents
8. Ability to recognize the literary devices used in a passage and to determine its tone and mood

9. Ability to determine a writer's purpose, intent, and point of view, i.e., to draw inferences about a writer

Once the classification of the numerous specific skills into the nine broader categories had been made, test items were constructed for each of the categories. Davis recognized that the first skill, knowledge of word meanings, is basic to measurement of others, and that some overlapping among the others is inevitable. In assembling the items into a test, he attempted to include the proportion of items testing each of the last eight skills that he judged to conform to the judgments of authorities on reading. Whether or not the skills isolated by this type of analysis are in fact independent and separately measurable is a question for statistical analysis.

It should be noted that the experience gained in the writing of the test items often contributes to a further clarification of the test objectives, and that frequently it becomes desirable to alter the test outline as the item-writing proceeds. There are few activities, in fact, that contribute more effectively to the clarification of objectives than the task of translating them into specific test behavior situations and the process of modifying and selecting items on the basis of the data secured from the experimental tryout. In general, therefore, it is undesirable to view the original test outline as "frozen." Rather, the test author should deliberately strive to improve the table of specifications as the test is being built. Indeed, the best time to write the final table of specifications for the test is after the items have been written and tried out and the final test assembled.

Analysis of curriculums and textbooks

As indicated before, the outline of a test intended to measure general achievement in a particular course of instruction should specify the test content in considerable detail. Moreover, each element of the content should be weighted roughly in proportion to its judged importance. When a teacher is constructing a test for use only in her own classes, perhaps she can be the sole judge of the appropriate weight for each subject-matter topic. Her judgment is quite likely to be influenced, whether consciously or unconsciously, by the relative emphases placed upon various topics in the textbooks she uses or consults. When a test is expected to have wide usage, perhaps nation-wide, it becomes essential that the judgments of a number of experts in the subject-matter field be solicited. Consequently, it has become standard procedure in the planning of a subject-matter test, especially one for a broad market, to analyze ten or a dozen of the more widely used textbooks in the field in order to secure a tentative list of topics to be tested and some indication of the appropriate emphasis to be given to each of them. The median number of pages de-

voted to a topic is often taken as a rough index of its importance, as judged by textbook writers and editors. This serves also as a crude index of the amount of time given to each topic by classroom teachers.

Analysis of curriculums from representative cities and counties scattered throughout the United States provides another method of determining the behavior objectives and subject-matter content to be measured by objective tests. Data obtained from curriculums provide a somewhat more accurate indication of the amount of time actually devoted in the classroom to each topic in a given field than do those obtained by analyzing textbooks.

Pooled judgments of authorities

In the construction of educational achievement tests, the advice and assistance of authorities should almost always be sought with respect to the topics of the test outline and the emphasis on each. The panel of expert consultants should not only command wide respect; but it should also constitute an adequate representation of professional thought in the field. For example, in the construction of a set of reading tests, it would be unwise to consult only with men known to be primarily concerned with the mechanics of eye movements and their training. A test that would satisfy this group of experts might be quite unsatisfactory to others in the field.

Ordinarily, subject-matter experts serving as consultants cannot be expected to write test outlines from "scratch." The most profitable way to use consultants is usually to submit to them a tentative outline, perhaps one based on analyses of the objectives of instruction as stated in representative curriculums and of subject-matter content as indicated by up-to-date and widely used textbooks. This suggestion is made on the assumption that the person initiating the test is either himself able to formulate such an outline or has an assistant with the necessary familiarity with the subject matter. The consultants are asked to criticize or revise the topics in the outline, to suggest additional topics, and to weight each of the topics finally agreed upon. Sometimes the weights derived from curriculum and textbook analyses and the test constructor's subjective judgment (if he has sound basis for such judgment) are presented on the outline given to the experts for criticism, but the disadvantage is that the experts tend to accept such weights and to refrain from suggesting changes in them. In general, therefore, a test outline without any indication of the emphasis to be given the topics is preferable for this purpose. A request to assign weights to the topics is then most likely to stimulate active and thoughtful consideration of the problem.

Determination of Appropriate Weights

An illustration of the use of expert judgment in constructing a test outline is provided by a paper-and-pencil mechanical comprehension test. Careful study of handbooks and manuals for mechanics suggested that eight topics for which paper-and-pencil test items could be constructed covered the field rather thoroughly. A list of these eight topics was, therefore, prepared and sent to sixteen individuals known to have a high degree of competence in mechanical comprehension or its measurement. Instructions accompanying the list were, in part, as follows:

> On the accompanying sheet are brief descriptions of each of eight topics. Please read the descriptions and then estimate what per cent of the total test should be devoted to each topic. Write your estimate in the parentheses given. . . . Your eight estimates should add up to 100 per cent. If you think some element of mechanical comprehension other than the eight listed is important in mechanical comprehension as a whole, please write a brief description on the line below the description of Topic 8.

In such instructions, provision might also be made for the judges to add topics that they believe are not covered by those given and to indicate appropriate percentage weights for them, again in such a way that all weights assigned sum to 100 percent.

The median percent assigned by each consultant to each topic was used as an indication of the contribution that each topic should make to the total variance of a test of mechanical comprehension. The eight topics and their weights were as follows:

Topic	Median Percent
1. Technical vocabulary	15
2. Tools	10
3. Hydraulics	7
4. Properties of materials and structures	9
5. Uses of mechanical devices	15
6. Mechanics	15
7. Electricity	8
8. Mechanical movements	21
Total	100

Note that the consultants were not asked to indicate the number of items of each type to include in a 100-item test of mechanical comprehension. Instead, they were asked to answer a question that is more easily answered by laymen; namely, "How much weight should each type of item have in determining the total score?" Again, it should be noted that for inexperienced test constructors, especially if they are not authorities in the subject-matter field in question, a more detailed outline of content would be desirable before actual construction of item writing proceeds.

Let us next consider how to translate the data into specifications regarding the number of items of each type needed to produce the desired weights.

It is sometimes assumed that the contribution of a part of a test to the total score is directly proportional to the number of items assigned to that part. One might casually suppose that in a test of thirty items it would necessarily follow that a twenty-item part would contribute twice as much to the total score as a ten-item part. This supposition is not necessarily correct, however, as can easily be shown. Let us first consider the relationship between the number of items in a test and its standard deviation. The variance of a test, expressed in terms of the variances and intercorrelations of its component items, can be written as follows:

$$\sigma_T^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 + 2(\sigma_1\sigma_2 r_{12} + \sigma_1\sigma_3 r_{13} + \cdots + \sigma_{n-1}\sigma_n r_{n-1,n}). \quad (1)$$

If the intercorrelations of the items were all unity and their variances all alike (that is, if the items were all equally difficult), equation (1) indicates that the standard deviation of scores on the test would vary directly in proportion to the number of items included in it. If the intercorrelations of the items were all zero and their variances all alike, equation (1) indicates that the standard deviation of the scores would vary directly in proportion to the square root of the number of items included in it. Both sets of assumptions are untenable in practice. In actual practice, the item intercorrelations are likely to be closer to zero than to unity; hence, it may be serviceable as a first approximation for the test constructor to figure that the standard deviation of a part of a test will increase a little more than in direct proportion to the square root of the number of items in it; that is, doubling the number of items in a part will make the standard deviation of a part about one and one-half times as large.

This is, however, a very rough approximation, since so much depends on the difficulty of the items added. For example, if the items added are all either very easy or very difficult, doubling the number of items might have only a slight effect on the standard deviation of the score distribution.

The second relationship we should consider is that of the contribution of a part of a test to the variance of the entire test. This is given precisely by a modified form of equation (1), as follows:

$$\begin{aligned}\sigma_T^2 = {} & \sigma_1(\sigma_1 + \sigma_2 r_{12} + \cdots + \sigma_n r_{1n}) \\ & + \sigma_2(\sigma_1 r_{12} + \sigma_2 + \cdots + \sigma_n r_{2n}) \\ & + \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ & + \sigma_n(\sigma_1 r_{1n} + \sigma_2 r_{2n} + \cdots + \sigma_n). \end{aligned} \quad (2)$$

Each line on the right-hand side of equation (2) constitutes the contribution of one part to the variance of the entire test. Inspection shows that this contribution depends on the size of the standard deviation of the part and on its correlations with all other parts. A test constructor does not have available the necessary data to solve this equation at the planning stage. It is common practice, however, to make a very crude estimate of the probable contribution of each part of a test to the total variance by assuming that the intercorrelations of the parts are all zero. On this assumption, the contribution of each part becomes directly proportional to the square of its standard deviation. Since the square root of the number of items in a part is very roughly proportional to the standard deviation of the part and since the latter may be taken as a crude approximation of the square root of the contribution of the part to the entire variance, it follows that the test constructor, by an extremely rough approximation, may make the number of items in a part proportional to the desired contribution of that part to the entire variance. Thus, for the test of mechanical comprehension for which percentage weights determined by pooled judgments of experts were presented earlier (page 167), the number of items on each topic to be included in a 100-item test would correspond very roughly to the percentage weights. It is important to emphasize the crudity of this approximation. If the test author has reason to believe that certain parts correlate highly with one or more of the others, consideration of the relationships in the equation above might lead him to depress the weights of such tests somewhat. Tryout of the test may reveal unexpected trends in the standard deviations of the parts and thus call for readjustment of the weights. It is only through the use in equation (1) of empirical data secured from a representative tryout that one can determine with high dependability the true weights of the various parts of the test or of the various types of items.

It should now be more clearly apparent why the statement was earlier made that it is not necessary or desirable to attempt to adhere rigorously to the weights or distribution of items in the various categories in the original table of specifications. In the first place, the weights themselves represent a series of compromises among the authorities consulted, and seldom correspond closely to the weights suggested or preferred by any individual authority. The weights are thus merely estimates of the "best" set of weights, and can seldom be regarded as highly reliable estimates. In the second place, the item writers may prove incapable of building convincing items for some categories in the numbers specified, but may produce a surplus of very good items in other categories. This sometimes may suggest that the former categories were originally none too well con-

ceived, or that they are concerned with unmeasurable or unteachable outcomes or traits, in which case the original weights might be somewhat reduced for the former and increased for the latter categories. Again, the result of the tryout may indicate that the authorities were somewhat unrealistic concerning actual achievements of the pupils to be tested, or may suggest an earlier or later grade placement of some of the outcomes tested. Finally, it must be recognized that the number of items assigned to the categories is only a very rough indication at best of the weights actually given to these categories in the total test score.

The foregoing does not imply that one may be careless about the distribution of items, or that careful planning is not worth while. On the contrary, it is highly desirable to prepare a table of specifications with the utmost care, and departures from the table and from its weights should be allowed only if good reasons for doing so present themselves. Such reasons, however, will often be encountered, and it would be unwise to disregard them or to assume that nothing may be learned from the tryout of the items that will lead to improvements in the table of specifications.

Determining Test Length

The number of items that should be included in the final form of a test is, from an idealistic standpoint, determined by the purpose of the test or the uses to which it is to be put, and by the statistical characteristics of the items. If important decisions regarding individuals are to be based on the test, it must be more reliable and hence must contain more items than would be the case if it were to be used only for crude group comparisons. Again, if the items of the test are very homogeneous in nature, the optimal number is lower than for highly heterogeneous item content or form. A test for which several part scores are to be obtained will require more items in order to produce a given degree of reliability of each part score than will a test on which only a single over-all score of the same degree of reliability is needed. These are all important considerations in setting test length.

Another factor influencing test length is the amount of time likely to be available or convenient for the administration of the test. Often, and particularly for educational testing, there is likely to be a maximum length of testing period related to administrative factors such as the usual length of a class period. Disregard of the exigencies of the practical situation may mean that an otherwise superior test will remain unused.

Given the allotted time, the test constructor, on the basis of experience with the type of material contained in the test, estimates the number of items likely to be appropriate. The number of items feasible for a particu-

lar period of time varies both with the form and content of the items and with the type of behavior they attempt to elicit. More specifically, the mere length of each item, insofar as it is related to reading time, will affect the number of items that can be used fruitfully. The vocabulary level of an item and the difficulty of its sentence structure bear directly on its reading time. Some mathematical items require much more time than others for computation. Items that tap the higher thought process take more time than those dependent on rote memory.

On the basis of the foregoing factors, together with the proportion of the total test that should be devoted to each type of item (as established in the test outline), the test constructor decides how many of each type of item the test should contain. Preferably, and especially if the test is to have widespread use, these numbers should be regarded as tentative until the test has been tried out. After the initial tryout, available data on reliability or validity can be used in modifying the original numbers of items of each type or the length of the total test. Ways to determine experimentally the optimum length of a test are described in chapter 10 (pages 336–38).

Determining the Number of Items To Be Constructed for Tryout

The number of items which should be constructed for tryout is always considerably larger than the number needed for the finished test. Review of test items, subsequent to tryout, by technical and subject-matter consultants on the basis of the statistical analysis of the tryout data commonly brings to light a number of items with such serious shortcomings that they cannot be salvaged. The planning of the number of items to be constructed should take cognizance of the anticipated item mortality. The loss will vary, of course, with the experience and competence of the item writer. It will also depend upon the type of item. Fewer surplus items are commonly needed, for example, on arithmetic fundamentals and on grammatical usage than on preferred diction. With some types of items, two or three times as many items as will eventually be needed should be constructed; in other cases the mortality may be only 10 to 20 percent. These problems are considered in detail in the chapters on tryout of the test, on analyses of the tryout data, and revision after tryout (chapters 8 and 9).

The number of items appropriate for inclusion in a test also depends on the extent to which it is intended to place a premium on speed of performance. A common tendency among test constructors is that of introducing a speed factor into tests originally planned as measures of level of

performance. Most tests of comprehension in reading, for example, yield scores that are a mixture of speed and level of comprehension. This matter should be given careful consideration when a test is planned, and the number of items included in the final form of the test should be dependent on a conscious decision to introduce speed of response or to eliminate it for practical purposes. (See chapter 9, pages 312–15.) The proposal made many years ago by Flanagan to obtain "power" scores by means of the repeating-scale technique should be given careful consideration when a test is planned (*4*). This technique has been used much less frequently than occasion for its use has arisen.

Determining the Time Limits for the Test

The problem of determining the amount of time in which the test is to be administered is ordinarily inseparable from that of determining the length in terms of number of items. Sometimes, as has already been suggested, the test is planned to fit a predetermined time interval, such as a 45-minute class period, in which case the problem is to estimate how many items will yield the maximum validity and reliability for that time interval. In other cases, the test is planned to yield a given minimum validity or reliability, in which case the problem is to estimate simultaneously both how many items and how much time will be needed to meet the standards set. This problem is considered fully in chapter 10, pages 336–38.

Planning the Types of Items

Before test construction is under way, and ordinarily at a fairly early stage in planning a test, decisions must be made concerning the types of items to be included in a test.

Relation to Scoring Facilities

A major factor affecting such decisions is the degree of scoring objectivity considered necessary or desirable. Tests intended for extensive usage are almost without exception made relatively objective; and among the various forms of so-called objective tests the more objective, such as the "multiple-choice," are to be preferred to the less objective, such as the "completion." This is true whether hand-scoring or machine-scoring is contemplated. If the tests are to be scored by machine, however, then selection of the highly objective types of items is clearly mandatory.

Relation to Administrative Facilities

In addition to the scoring facilities, the probable quality of the test administration has a bearing on the types of items to be selected. If the test is likely to be given by different persons, and by persons with little

training or experience in test administration, then it behooves the test planner to take precautionary measures to minimize the difficulty of administering the test. Inclusion in the test of items of only one type greatly simplifies the administration of the test, especially when a single over-all time limit applies. Use of a single item form reduces the amount of time necessary for giving directions and can help to eliminate the possibility that failure to understand directions may invalidate test scores.

Use of Few Versus Many Types of Items

Some have argued that use of several different kinds of items lends interest to a test through its variety. However, the interest value of a test depends primarily upon the quality of the items, rather than upon their external form, and if the items are competently constructed, even though all of one form, there will usually be no problem of maintaining interest.

The form of items used must be suitable in relation to test content. Some test constructors believe (and with some justification) that certain types of content or particular topics lend themselves more readily to particular item types than to others. It may even be felt that there should be sufficient flexibility in the selection of item forms that the particular item topic is allowed to determine the form, which hence should not be specified in advance. Probably most specialists in the field of test construction would regard this as an undesirable extreme of a position that in moderation is reasonably acceptable.

On the whole, arguments for use of as few forms as possible, and preferably only one, for a given test or area of subject matter probably outweigh arguments for variety.

Achieving Appropriate Item Difficulty

Relation to Test Purpose

Test planners should give careful attention to the desired level and range of difficulty of test items and discuss this topic with the persons who are to write the items. Technical considerations pertinent to what constitutes the optimal level and range of difficulty in relation to the purpose of the test are treated in chapter 9. Once it is decided what level and distribution are to be sought, the task is to attempt to construct items that will yield the desired characteristics.

Factors Affecting Item Difficulty

As is made clear in chapter 9, an index of item difficulty may be obtained by estimating the percentage of examinees who know the answer to it.

Primary factors affecting the difficulty of an item are the nature of its content and the type of behavior it requires of the examinee. A test constructor may quite readily judge that an achievement test item on a topic to which eighth-graders have not been introduced will be very difficult for them or that an item requiring a new application of a scientific principle will produce more failures than one based on simple recall of the principle. Such judgments are likely to be helpful.

Apart from intrinsic content, however, other factors influence item difficulty, often in very subtle ways. Unusual vocabulary may markedly, albeit unintentionally, influence responses to an item. Awkward sentence structure and undue formality in the style of the language used often have unpredicted effects. A shift from use of third person to first or second may make an item significantly easier. Even such apparently extraneous factors as the form of the item and the directions to the examinees may affect item difficulty. Matters such as these should be brought to the attention of the item writers, in addition to the more obvious considerations related to the intrinsic nature of the subject matter and type of behavior involved.

Using Judgments and Statistical Indices of Difficulty

The use of subjective judgment in estimating item difficulty at the stage of item construction is to be encouraged. Such judgments, when based on all available experience, are distinctly helpful in leading to the construction of items of the desired difficulty. They are not ordinarily sufficiently accurate, however, that maximum efficiency of the finished test can be secured on this basis alone. Except in rare situations, the items must be tried out on a sample representative of the universe for which the final test is intended. As explained in chapters 8 and 9, selection of items for the final form of a test should systematically take into account the difficulty level of the items. In order to secure anything like maximum efficiency of measurement, it is necessary to adjust the distribution of difficulty of the items in a test so that it will be appropriate to the subjects being examined and to the purpose or purposes of the examiner. This final adjustment is accomplished by using difficulty indices computed on the basis of data obtained during the tryout of the test.

Sometimes the tryout data reveal that not enough items of a given level of difficulty were constructed. The items may prove, in general, to be too difficult or too easy for the subjects in the experimental group. The data may show that an efficient measuring instrument cannot be constructed by using only the items that were tried out. Careful attention to the most appropriate range of difficulty while the items are being written and before the tryout form is assembled helps to preclude this possibility. All items

should be keyed and criticized by subject-matter experts and test technicians. Items that are apparently outside the range of difficulty that will be required should be discarded. If there seem to be too many items at one level of difficulty and not enough at another level, that condition should be corrected before the tryout form is prepared for administration. Careful planning and editing of the experimental form of a test can save a great deal of money and trouble and permit construction of tests that approach maximum efficiency. Unfortunately, the painstaking care and psychological insight needed to produce tests of high quality and high efficiency are often lacking in the process of test construction, perhaps because they cannot be routinized as can item analysis techniques and other statistical procedures.

In large test construction agencies, a test planner sometimes has available a large reservoir of items whose difficulty is known from previous tryouts. In such a situation, the test outline can advantageously include the plan for the range and distribution of difficulty indices for each subject-matter area or each type of behavior included in the outline. Then the selection of items can be done systematically to insure not only the desired content but also the desired difficulty characteristics.

Planning Other Test Development Operations

In the introduction to this chapter it was stated that planning is essential throughout a test construction project. Once one has defined the purpose of a test, prepared an outline of its content, selected the type or types of items to use, and taken precautionary measures to insure appropriate item and test difficulty, he can still avoid countless mishaps by anticipating them. This section, covering practically all of the remaining steps in a typical test development project, points to additional problems to which the test planner should give early attention.

Construction of Items

The question of who is to construct the items needed for a particular test warrants attention. In some instances, where the subject matter concerned is fairly simple, the most feasible plan is to have the items constructed by a specialist in test construction who either knows or can readily acquire the requisite subject-matter knowledge. If test technicians unfamiliar with a more complex field of knowledge or skill attempt to construct items in that field, however, the test as a whole is very likely to be faulty. Some of the individual items may not have a defensible answer or may have more than one defensible answer. Perhaps more serious, the items are likely to depend to too great an extent on material readily available in

books, and too little on interpretations that would evidence real understanding. In other words, the resulting tests are likely to be open to the charge of being "textbookish." As the subject-matter field becomes more complex and specialized, then, it becomes more and more important that the item constructor have had extensive training and experience in the field.

In some situations, there may be available for item writing persons who not only know the content area thoroughly but who also have had training and experience in the specialized techniques of item writing. This is probably the exceptional case, however. It is ordinarily necessary to provide the subject-matter specialist with specific training and supervision in item writing. The necessity for providing such training again calls for decisions on such matters as how much training is essential, the proper timing of various aspects of the training, the extent to which it can be conducted in group sessions, the amount of time that can profitably be devoted to discussion and joint revision of individual items, and so on.

The rate of item construction is another matter on which planning is required. No single rate of construction applies to all types of items or to all item constructors. Some item forms require more time than others. True-false items can be written at a more rapid rate than multiple-choice items, for example. There are also wide differences attributable to the subject-matter area: vocabulary or arithmetic items, for example, may be constructed by an experienced item writer at the rate of, say, fifty or more a day, while a rate of from three to ten a day may be entirely acceptable for the production of items for a test of ability to interpret reading materials at an advanced level. For a given item form and area of subject matter, furthermore, some item writers are capable of producing items several times as rapidly as others.

The deadline for completion of the test must be taken into account together with the expected rate of item production in deciding how many item constructors are needed. Sometimes the deadline is such that a number of item constructors must be used, even though this adds to the problem of training and creates need for coordination of their work.

Apart from the urgency of test deadlines, there is often a further reason for having several persons construct items for a test. Even with the use of a carefully detailed outline of test content, a set of test items constructed by one person is likely to be influenced by whatever biases he may have in the content area; and such biases may escape the later reviewers or be difficult to correct. Use of several item constructors with different backgrounds of training and experience leads to the reflection in the test of somewhat different points of view or emphases within the subject-matter area.

There is also need for planning in order to be sure that any source materials needed can be available at the time they are needed. This may seem an obvious point, but it is one that is too frequently overlooked.

Review of Items

As a general rule, test items should be reviewed, before tryout on any sizable number of subjects, from three points of view: the accuracy and appropriateness of their subject-matter content, their technical merits apart from content, and their "editorial" quality. Effort spent in intensive review of these three types will obviate the need to try out unusable items and will improve the general quality of the items tried out.

Sometimes one person has skill in all three areas and thus is able to combine the three types of review. More frequently, different persons are used for the three quite different purposes. No hard and fast rules can be set down as to how many reviewers for each function should optimally be used. The numbers depend to a considerable extent on the type of items and the particular subject-matter area and also on the persons involved. Individuals who earn their livelihoods as test constructors, for example, should develop skill in handling such editorial matters as punctuation, spelling, diction, uniformity of style, and so on, so that a review by only one additional person from an editorial standpoint should suffice. Perhaps two test technicians might profitably review the items from the technical point of view. Ordinarily a larger number of persons should review the subject-matter content intensively—somewhere between, say, three to ten, depending on the complexity of the area. Often it is desirable to bring the subject-matter reviewers together to discuss points of difference and attempt to achieve a final version of each item that meets all objections.

Records on Each Item

Before item construction is begun, the test planner must decide what records are to be kept on each item. If the first drafts of items are to be retained, even if only until a test is in print, having them all on sheets of paper or cards of a single size will facilitate handling and filing. If reviewers' comments are to be retained, provision for the comments to be recorded in a uniform way will be helpful. If a record of any written source used in construction of the item is desired, a plan for this record should be made before the item reaches its first draft; otherwise the book may have been called back to the library. If a record of who constructed each item is desired, in case different persons are working on a test, his identity should be recorded on the initial draft and later transcribed to whatever final record form is maintained for the item. Plans also need to be made for

keeping a record of item use and whatever statistical data are desired for each item. More detailed treatment of item record systems is given in chapter 9.

Tryout and Analysis of the Test Items

Chapter 8, devoted to the tryout of tests, makes clear that many questions about tests can be answered with assurance only on the basis of tryouts. The topic is introduced here because only through attention to planning for the tryout and analysis of test results can there be any assurance that the tryout will be adequate. The purposes of the tryout should be outlined in some detail, and for each the question asked whether or not the tryout is being planned in such a way as to serve the purpose.

One feature of the tryout that warrants expert consideration is the selection of the sample on which the tryout is to be conducted. This attention to the sample should cover not only whether the sample is reasonably representative of the examinee group from the standpoint of abilities but also whether its members are likely to have or can be made to have about the same degree of motivation as the examinees.

Another feature of the tryout that should be mentioned here relates to timing. The test planner must be sure to allow ample time for administering and scoring the tryout form, for analyzing the results statistically, and for processing whatever revisions are indicated by the tryout. This planning requires attention both to manpower resources and to deadlines for the final form of the test.

Compilation of the Items

A group of items is not necessarily a test. Once the individual items have been constructed, the problem remains of selecting from among those that survived the review process and the tryout those which are to constitute the test, and of arranging the selected items into an appropriate order.

One problem often arising in test compilation is that of avoiding undesirable overlapping among the items. There ordinarily should not be two items so closely related that an examinee who can answer one correctly automatically can answer the other (unless the test constructor may wish to weight the factor tested more heavily than the other items). Nor should one item contain a clue to the correct answer to another item. Such overlapping of item content reduces the reliability of the test. If an outline of test content has been prepared in great detail, the possibility of the first type of overlapping should be minimized; it still may occur, however. And no matter how specific the test outline, the second type of overlap is probable, particularly if several persons working independently have constructed the

items and if all items have not been reviewed within a relatively short period of time.

The test compiler must give attention to the adequacy of the coverage of the subject-matter field by the selected items. Again, to the extent that the test outline is comprehensive, this problem will be reduced. But if the outline presents only a few broad categories, a great share of the responsibility for insuring that the test will serve its purpose rests with the compiler.

In the assembling of items into a test, various questions relating to the order of the items must be considered. Occasionally some of these questions are anticipated in the preparation of the test outline, a special form of which may be prepared to show the order in which the several divisions of subject matter are to be tested.

Various ways of grouping items have been tried. The nature of the items and the character of administrative conditions usually are the determining factors in deciding how to group items. If the items are homogeneous in content and in difficulty, so that they are essentially interchangeable, the order of the items is inconsequential.

If a test of heterogeneous content is to yield only one score, the items of each type can be grouped together in subtests and a separate time limit provided for each group of items or subtest so as to insure that each type of item receives its proper share of attention from the examinees. Ordinarily the items within each subtest would be arranged in order of ascending difficulty. The order of the subtests might be based on their relation to a logical organization of the subject matter. Perhaps, for example, the broader or more general areas should come first, followed by the more specialized areas.

If many examinees cannot finish the entire test in the time limit, and it is impracticable to make use of a separate time limit for each group of items, it may be desirable to present first the easiest items of type 1, followed by the easiest items of type 2, etc., until the easiest items of all types have been presented. There will then follow a set of moderately easy items of each type, ending with the most difficult item of each type. This is the so-called spiral omnibus arrangement. One of its disadvantages is that the sophisticated subject is likely to skip over the segments he thinks he cannot do quickly or accurately, while the more conscientious or less sophisticated but equally apt subject plods along, taking the items as they come. Naturally, the sophisticated subject is likely to get a higher score, which means that the validity of the test suffers because of the introduction of unwanted variance pertaining to personality characteristics and experience with tests. What is worse, from one point of view, is that the sophis-

ticated subject realizes what he has done and knows that he has "got away with it." Consequently, his opinion of objective tests is likely to be mildly contemptuous, to say the least.

The spiral omnibus arrangement has other disadvantages, too. Among these are the fact that the examinees have to keep changing their mental-set from one type of item to another. It may be uncommon for this sort of performance to be required in real life situations. If, however, one's purpose is to attempt to test "mental flexibility," a spiral omnibus test might be useful. Needless to say, there is no way to obtain satisfactory part scores from a spiral omnibus test for research purposes. Part scores derived from this sort of test must ordinarily be regarded as expedients that are useful for some practical purposes. A test constructor ought to be sure that the use of separate time limits for different parts of a test or use of the repeating-scale technique is impracticable before he decides to forego the use of groups of more or less homogeneous items.

Directions to Examinees

It is usually desirable to prepare the directions to the examinees for the finished test, in at least rough form, before the items are constructed. Early preparation of these directions will focus attention of the item writer on the problem of adapting the items and item forms to the background and experience of the examinees, and will lead to clearer thinking about test length and time limits. Certainly these directions should be prepared in advance of the tryout, since it is very important that the same directions be used in the tryout as in the finished test. The preparation of these directions is considered at length in chapter 10, pages 351–65.

Review of the Test as a Whole

Once the proven test items have been assembled, the time limits established, and the instructions to the examinees prepared, the test should again be reviewed from three points of view: the technical, with particular attention to principles of measurement, including those relating to item form; the subject matter, with attention to appropriateness of content and especially to the accuracy of the scoring key; and the editorial, this time with special attention to appropriate over-all format, to editorial consistency from one item to another, and to the proper relation between the instructions to examinees and the test itself.

As in the case of the review of individual items, one person may possess competency in more than one of these fields. Ordinarily, however, the test planner will have to arrange for different persons to consider the test from the different points of view. The number of such reviewers and the inten-

sity of their review will depend in large part on the nature of the content and the excellence of the previous reviews.

Reproduction of the Test

The way in which a test is to be reproduced should ordinarily be given some consideration even before item writing begins. The method of reproduction to be used sometimes determines what types of items may be employed. For example, if the test is to be lithoprinted, liberal use of photographs, charts, tables, maps, etc., will not materially affect the cost of the test, and their use may be especially encouraged. If the test is to be reproduced by letterpress printing or by Mimeograph, however, use of visual materials will have to be curtailed for reasons of cost and other practical considerations.

The method of reproduction to be used will also affect the preparation of the final copy. If a test is to be reproduced by a photo-offset process, for example, the final copy must be exactly as it is to be reproduced (except, of course, for size, since reduction or enlargement is possible). This means that more expert typing is required than for a test to be set in type, for in the latter case interlineations or other types of corrections that the printer is to make can be indicated directly on the copy. If, on the other hand, a test is to be reproduced by Mimeograph or Ditto or a similar process, as might be the case for a test for use with a single class, then the final copy often can be made from item cards directly onto stencils.

The test planner must give preliminary consideration to any drafting work that the final test will require and govern his plans according to the type of reproduction of the test that appears most practical in the particular situation. For example, if all test pages are to be multilithed, the test planner must decide whether all drawings are to be drawn in the size in which they are to appear or whether they are to be reduced before being fitted into place on the page or whether the entire page is to be reduced.

Many of the details of test reproduction to which the test planner will need to give attention are given more complete consideration in chapter 11.

Scoring the Test

Before test planning can progress very far, a decision has to be reached as to whether the scoring is to be subjective or objective. Since this book is addressed primarily to the question of developing objective tests, no further consideration will be given to subjective tests except to note that they are much more difficult to score and require even more careful planning than objective tests in attempts to reduce scorer unreliability.

One of the first questions regarding the scoring of objective tests that

arises is whether the scoring is to be by hand or by machine. The answer will depend partly on the availability of hand-scorers as against the availability of scoring-machine facilities and partly on the predicted costs of the two types of scoring. It is clear that there is a limit to the number of tests to be scored below which machine-scoring is not economical. In other cases, where the number of tests to be scored is very large, it is probably a moot question whether machine-scoring is cheaper than hand-scoring, although many persons believe this is true. In some instances, good design of answer sheets and intensive training and high motivation of scoring clerks may make hand-scoring more economical than machine-scoring. For any sizable testing program, this question is worthy of careful scrutiny before a decision is reached.

Occasionally a test constructor may believe that the type of content with which he is dealing does not lend itself readily to the multiple-choice form demanded for machine-scoring. In such a case it may be decided that the nature of the material rather than the possibility of machine-scoring should determine the item form to be selected.

Whether a test is to be scored by machine or by hand, use of a separate answer sheet is usually advantageous from the standpoint of scoring economy alone. Another factor that has bearing on the use of separate answer sheets is the type of test material or the nature of the ability being tested. If, for example, performance on a test is known to depend largely on perceptual speed, use of a separate answer sheet might introduce additional factors that would have unintended effects on test scores.

If hand-scoring is to be used, the type of scoring key to be applied must be determined. There must also be decisions on whether right answers or wrong answers are to be marked or whether the scorer is simply to record a score without marking the answers. The stage in the scoring at which any weighing and combining of subtest scores are to be accomplished must not be neglected.

The test planner, if he is also responsible for scoring a large number of tests, must plan how many scorers can be used most effectively, how they should be selected, and what training and supervision they will require. He must also plan how the scores are to be checked. The whole problem of scoring the test is thoroughly discussed in chapter 10, pages 365–413.

Construction of Alternate Forms of a Test

If it can be anticipated that several forms of a test are to be needed, plans for their construction should be made when the construction of the first form is being planned.

The test planner must determine what definition of comparability he is going to adopt or what he is going to accept as the minimum essentials for comparability. He would ordinarily want to give attention to whether the two forms test the same functions and whether they yield score distributions of the same type and with the same central tendency and dispersion (*1*). He would need to decide to what extent he could try out the test in a preliminary edition to determine whether these conditions were met or what adjustments would be desirable.

Regardless of the opportunity for test tryout, it may be said that tests intended to be comparable should be based on the same outline of content. The more detailed the breakdown of subject matter in the outline, the more it facilitates construction of comparable forms. If the outline is so detailed that it indicates the content and behavior to be tested by each item and perhaps even the difficulty of the item, then two tests constructed on the basis of the outline will contain pairs of items that are more or less interchangeable insofar as the judgment of the item constructor is concerned.

Construction of comparable forms by attempting to construct pairs of items of identical difficulty and testing identical abilities raises the question of just how similar the items within each pair should be. Without an extended discussion of this question, the best answer that can be given is that the members of the pair should not be so related that knowledge of the answer to one would immediately tell a person the answer to the other, but they should be so similar that in general a person able to answer one is likely to be able to answer the other.

Before tryout, the decision on comparability calls for skilled judgment on the part of the test constructor. There is always the danger that the items so constructed will be so similar that exposure to the one will be of assistance in answering the other. This is not important if the same persons are not going to take the same forms; but if this were not likely, there would be little necessity for more than one form in the first place.

Conclusion

This chapter is best ended with a restatement of its purpose. In a very real sense this entire book is a book on test planning. Many of the topics that have been considered in this chapter will be taken up again in much higher detail in later chapters. This chapter, then, has merely provided an introduction to, or a preview of, what follows. It has been intended primarily to orient the reader, to help him better to view all phases of test construction in their proper relation to one another, to recognize the need for careful coordination of the many separate phases, and to appreciate

the great importance of anticipating all subsequent steps in test construction as each next step is taken. It may serve some of these purposes better if it is read both before and after the other chapters are read.

Selected References

1. ADKINS, DOROTHY C., *et al. Construction and Analysis of Achievement Tests.* Washington: Government Printing Office, pp. 202–6, 1947.
2. ASHFORD, T. A. "The College Chemistry Test in the Armed Forces Institute," *Journal of Chemical Education,* **21**: 386–92, 1944.
3. DAVIS, FREDERICK B. "Fundamental Factors of Comprehension in Reading," *Psychometrika,* **9**: 186, 1944.
4. FLANAGAN, J. C. "A Proposed Procedure for Increasing the Efficiency of Objective Tests," *Journal of Educational Psychology,* **18**: 17–21, 1937.
5. VAUGHN, K. W. "The Interpretation and Use of the Professional Aptitude Test," *Graduate Record Office Bulletin,* 1–16, 1947.

7. Writing the Test Item

By Robert L. Ebel
State University of Iowa

Collaborators: Dorothy C. Adkins, *University of North Carolina;* Louise Witmer Cureton, *University of Tennessee;* Charlotte Croon Davis, *Cooperative Test Service* (formerly); Frederick B. Davis, *Hunter College;* W. W. Turnbull, *Educational Testing Service*

Any test consists of a number of tasks to be performed by the examinee. Some of these tasks are scored as indivisible units. Others are subdivided for scoring purposes.

An "item" may be defined as a scoring unit. An "exercise" may be defined as a collection of items that are structurally related. For example, a matching exercise may consist of five items. A reading passage and the items based upon it also constitute a test exercise. An "objective" item or exercise is one that can be scored by mechanical devices or by clerks who have no special competence in the field.

The present discussion of item writing is directed primarily toward objective items and exercises used in paper-and-pencil tests of educational achievement. However, many of the suggestions made in this chapter will apply to other types of paper-and-pencil tests.

Item writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice. It demands, and tends to develop, high standards of quality and a sense of pride in craftsmanship.

Item writing is essentially creative. Each item as it is being written presents new problems and new opportunities. Just as there can be no set formulas for producing a good story or a good painting, so there can be no set of rules that will *guarantee* the production of good test items. Principles can be established and suggestions offered, but it is the item writer's judgment in the application (and occasional disregard) of these principles and suggestions that determines whether good items or mediocre ones will be produced.

Those who have not tried to write objective test items to meet exacting standards of quality sometimes fail to appreciate how difficult it is to write such items. The amount of time that competent item writers devote to the

task provides one indication of its difficulty. Adkins has pointed out that experienced professional item writers regard an output of five to fifteen good achievement test items per day as a satisfactory performance.[1] This contrasts sharply with the widely held notion that any good instructor can produce an acceptable test in an evening or two. Further evidence concerning the difficulty of item writing is provided by the amount of money that critical test producers appropriate for the work. It is not uncommon for item writers to receive \$2.00 or more for the production of a single good test item, or to be paid at the rate of from \$2.00 to \$4.00 per hour. The cost of a finally reviewed and approved test item may run from perhaps \$3.00 to \$10.00, depending on the content involved and the care exercised.

Extensive use of statistical methods for the analysis of responses to test items has seemed to imply that test production can be made a statistical science in which the skill of the item writer is of secondary importance. This notion is based on a misconception of the role of item analysis. Item analysis data do indeed often call attention to specific weaknesses within otherwise good items, and thus provide clues by which an ingenious item writer can make improvements. Such analyses may also make possible elimination of some of the weak items from a group, but, under usual circumstances, the amount of improvement that can be effected by this process is slight. The appropriate role of item analysis is fully discussed in chapter 9. For the present it is sufficient to observe that the analysis of test items in no way lessens the necessity for skill and care in the original writing of them.

Requirements for Writing Good Items

The combination of abilities required for successful writing of educational achievement test items can be specified easily in general terms. It is much more difficult actually to find persons who have these abilities.

First, the item writer must have thorough mastery of the subject matter being tested. The term "mastery of subject matter," as here used, has broad connotations. Not only must the item writer be acquainted with the facts and principles of the field, but he must be fully aware of their implications, which is to say that he must *understand* them. He should be aware of popular fallacies and misconceptions in the field. This is particularly necessary in the construction of items that provide suggested responses and therefore require incorrect responses which appear plausible to the poorer examinees.

[1] Of course, certain types of items are much easier to prepare than others. Ten acceptable vocabulary items could probably be written in the time often required to produce one satisfactory item intended to measure a complex understanding.

Sometimes test items are written on the basis of collaboration between test technicians and subject-matter experts. While this procedure is not ordinarily as productive of good items as is the work of a single person who fortunately combines test competence and subject-matter expertness, it yields far better items than would be produced by either specialist working alone. In practical test construction, collaboration of this sort is very helpful. The effectiveness of the collaboration depends not alone on the degree of competence of each specialist, but also upon the extent to which each shares a general background in the specialty of his partner.

Second, and of utmost importance, the writer who prepares items for use in tests of educational achievement must possess a rational and well-developed set of educational values (aims or objectives) which are not mere pedagogical ornaments, but which so permeate and direct his thinking that he tends continually to seek these values in all his educational efforts. (See chapter 5, pages 121–41.) It is difficult, if not impossible, for one whose sense of values is inadequate or inoperative to produce good achievement test items consistently. He may have at his disposal detailed syllabuses describing course content. He may be well acquainted with what goes on in typical classrooms where certain principles and abilities are being taught. But if he is not clearly aware of the educational values that should be directing the teaching and learning, he is almost certain to emphasize the superficial at the expense of the essential.

Third, the item writer must understand, psychologically and educationally, the individuals for whom the test is intended. He must be familiar enough with their probable levels of educational development to adjust the complexity and difficulty of his items appropriately, and to know what will constitute plausible distracters for multiple-choice items. He must have enough insight into their probable mental processes when confronted with various types of questions to avoid ambiguity, irrelevant clues to correct responses, or the measurement of extraneous abilities.

Fourth, the item writer must be a master of verbal communication. Not only must he know what words mean and insist on using them with precise meanings, but he must also be skilled in arranging them so that they communicate the desired meaning as simply as possible. Always, he must be critically aware of various possible interpretations which the examinee may make of the words and phrases in the item. It is probably true that no sentences are read with more critical attention to meanings, expressed and implied, than those which constitute test items.

Fifth, the item writer must be skilled in handling the special techniques of item writing. Obviously he needs to be familiar with the types and varie-

ties of test items and with their possibilities and limitations. Obviously he needs to know the general characteristics of good items and needs to be aware of the errors commonly made in item writing. But excellence in item writing demands more than this. It demands imagination and ingenuity in the invention of situations that require exactly the desired knowledge and abilities. It demands ability to identify the crucial element in each problem situation so that the corresponding item will be as direct and concise as possible. Most of all, it demands the skill and judgment that come only with experience. In test construction, as in other fields, the author must usually learn to write by writing.

In consideration of these requirements, several things should be clear. The process of constructing good test items is not simple. Not all individuals are equipped to master it easily. The abilities needed are too deeply rooted and too slow of growth to be produced in a short period of time. Manuals and rules may provide useful guides and helpful suggestions for item writing, but there are no automatic processes for the construction of good test items. Even an item writer who possesses the needed abilities will find that his success varies with the amount of energy and time which he is willing to devote to the task. Recognition of the skill and pains which go into the production of a good test is a prerequisite to improvement in item writing.

The high standards for item writing implied by this discussion have not been met generally in the past. A very large number of educational tests have been produced by item writers who lack the qualifications suggested, and whose tests consequently fall short of the standards set. This situation is likely to continue in some degree for an indefinite future period. Meanwhile, in the interests of better testing, it is desirable that ideal standards as well as present shortcomings and obstacles be recognized. Consistent efforts must be made to overcome the faults and approach the ideals. Not every item writer can hope to possess ideal qualifications, but each can improve his tests by rational application of the specific suggestions offered in the remainder of the chapter. Furthermore, those who direct test construction agencies must recognize more fully the paramount importance of the item writer in the whole process, and must offer rewards and facilities that will attract to the task of item writing the very highest level of intellectual competence.

Previous Studies of Item Writing

The problems of item writing have not received the attention they deserve in the literature on testing. There are abundant references on the controversial aspects of the testing movement, on the techniques of statisti-

cal analysis of test data, and on the applications of test scores in guidance, placement, and evaluation. But the basic problem of writing good items has been neglected. An unpublished survey recently made by the writer of nearly one hundred and fifty periodical references on the construction and use of objective tests revealed only five articles which dealt directly with problems of item writing.

Most textbooks on educational measurement show a similar lack of emphasis on item writing. Reading the chapter titles of these textbooks, one would never guess that *a* major problem, perhaps *the* major problem, facing the test specialist is that of writing good items. There are a few notable exceptions. The *Manual of Examination Methods* (*8*), produced by the University of Chicago Board of Examiners in 1933, contains some suggestions for item writing, although it is devoted mainly to illustrating various types of test items and the application of these types in various subject areas. In 1936 a committee sponsored by the American Council on Education produced a manual, *The Construction and Use of Achievement Examinations* (*3*), which contains specific suggestions for the writing of several widely used types of objective test items. More recently, Adkins and others prepared a volume on *Construction and Analysis of Achievement Tests* (*1*), which contains a section dealing with item writing. On the whole, however, the amount of reflection, discussion, and publication directed to the problem of writing objective test items has fallen far short of the requirements of the task.

Research on problems involved in item writing has likewise been scanty, particularly in recent years. This is due in part to recognition of the difficulties involved in such research. Many early studies were inconclusive or actually misleading. These faults have been recognized, but no generally satisfactory means of avoiding them has been discovered. It is extremely difficult even to identify the variables which affect the quality or functioning of a test item. It is even more difficult to control these variables in an experimental situation. As previously pointed out, the sense of values, the skills, and the ingenuity of the item writer play a large role in determining the quality of a test item. These characteristics are so complex and so diverse in manifestation that they are not easily subject to experimental control.

A further criticism of many early studies is that they were directed toward inconsequential problems. For example, one popular subject of research was the characteristic reliability and validity of various item forms. It is now clear that any such characteristic differences as may exist among item forms are of trivial consequence when compared with the extreme differences observed among items of the same form.

Thus, it is that few of the suggestions made in this chapter will be sup-

ported by concrete research findings. In the main, they represent distillations of the experience of those who have been successful in test construction. Such distillations of experience are the best available guides to effective item writing. It is to be hoped, however, that aggressive attempts will be made in the near future to verify at least some of the recommendations empirically.

Ideas for Test Items

The Nature of Item Ideas

Every test item begins with an idea in the mind of the item writer. The production and selection of ideas upon which test items may be based is one of the most difficult problems facing the item writer. While the test plan (see chapter 6) outlines the areas to be covered by the test and indicates the relative emphasis each area should receive, it does not ordinarily specify the content and purpose of each individual test item. The item writer is given the responsibility of producing ideas and developing them as items that will satisfy the specifications of the best plan. The quality of the test produced depends to a considerable extent upon the item writer's success in dealing with this problem.

If expressed formally, the idea for a test item would consist of a statement identifying some knowledge, understanding, ability, or characteristic reaction of the examinee. The following are examples of item ideas.

a) Knowledge of the steps in the enactment of a federal law.
b) Understanding of the relation of tides to the position of the moon.
c) Ability to add two terms to the first three of a geometrical progression.
d) Reaction to picketing by members of a union.
e) Ability to infer the meaning of a word from the context in which it is found.

In actual practice item ideas are seldom formally stated. Usually they exist only temporarily and with no verbal explicitness in the mind of the item writer. An idea occurs to him, he judges its probable contribution to the test, and, if this is satisfactory, he proceeds immediately to write it in approximately its final form. This procedure is quite adequate in practical test construction. There are few cases in which a useful purpose would be served by preliminary formal statement of the item idea. But in the present discussion it is useful to consider the essence of an item apart from the phrasing in which it is presented.

The Sources of Item Ideas

There is no automatic process for the production of item ideas. They must be invented or discovered, and in these processes chance ideas and

inspirations are very important. It is possible, however, to stimulate these processes by appropriate material. One source of stimuli is provided by instructional materials. Textbooks, course outlines, statements of objectives, lists of essential principles or basic abilities or frequent errors or common misunderstandings, discussion questions, and even questions from other tests are likely to suggest useful item ideas.

A second type of material which is likely to stimulate the production of item ideas is provided by the written work of the students themselves. Their expressions of ideas on issues and problems may reveal points of difficulty which can be made the basis of discriminating test items. A third source of item ideas is provided by job analysis. This procedure, borrowed from the constructors of selection tests for use in government, business, and industry, may also be applied in the construction of educational achievement tests. In using it, the item writer asks, "What is an individual who is proficient in this area expected to be able to do?" or, "How will the individual who is proficient in this area differ from one who lacks proficiency?" The answers to these questions may suggest valuable ideas for test items.

The difficulty of obtaining item ideas depends, of course, upon the nature of the items desired. If the purpose of a test is simply to determine whether or not the examinees possess certain information, the item writer needs only to consult various sources of this information and base items on some of the statements he finds there. The simplicity of this process is probably one of the chief reasons why educational achievement tests have been so often overloaded with informational items. If, on the other hand, a serious effort is being made to measure understanding, ability to interpret and evaluate materials, and similar characteristics of the examinee, the task of obtaining item ideas is much more difficult. In this case the item writer must acquire a thorough understanding of the subject and must work to invent appropriate novel situations.

The Selection of Item Ideas

The process of selecting item ideas goes on simultaneously with the process of inventing them. Skill in item writing depends not only upon prolific inventiveness but also upon discriminating judgment in the selection. In selecting item ideas, the writer must consider their appropriateness, importance, and probable discriminating ability.

Obviously the item ideas should be appropriate, which means that they must be in keeping with the test plan. They should deal *only* with those aspects of achievement that the test is intended to cover, but they should deal with *all* of those aspects. The number of items related to each aspect

of achievement should reflect the allocation of emphasis specified in the test plan.

Item ideas should also be selected on the basis of importance. While the test plan usually indicates the *general areas* of achievement that are regarded as important, it is the responsibility of the item writer to select important *specific aspects* of achievement as item topics.

The term "importance" as here used is almost synonymous with "usefulness" in its broadest sense. The ability to add is important because it is so frequently useful. The ability to apply artificial respiration is important because, once in a lifetime, it may be critically useful. The concepts of atomic structure are important because they are, within a limited area of human activity, fundamentally useful as a basis for understanding physical and chemical phenomena.

In contrast to these fundamental, crucial, or frequently used aspects of achievement are the superficial details, the incidental observations, and the explanatory illustrations and comments that have no enduring significance. For example, it is relatively unimportant that the birth year of Woodrow Wilson coincided with the first Republican presidential campaign. Yet many items dealing with such trivia have found their way into tests of educational achievement.

Finally, in the selection of item ideas it is necessary to consider their probable ability to discriminate between those who possess and those who lack a given understanding or ability. The chief purpose of most tests of educational achievement[2] is to rank examinees as accurately as possible in order of their attainments. Only those items that are answered correctly by the better-qualified examinees and missed by those who are not well qualified contribute to the effectiveness of such tests. While an item writer cannot be certain in advance how a given item will perform, his selection of item ideas should be guided by several general principles. Aspects of achievement that are thoroughly mastered by all examinees, or those that for various reasons are hardly mastered at all, are likely to provide few useful discriminations. Further, if the specific ability called for in an item is not highly related to general proficiency in the area covered by the test, the item will discriminate poorly. Of course, the discriminating ability of an item also depends upon how well it is written, but even the best of writing will not convert some ideas into discriminating test items.

Inclusion of poorly discriminating items is occasionally justifiable because of the contribution which the item makes to the apparent validity of the test, or because of its probable influence on study and teaching. For ex-

[2] The statements in this paragraph do not apply to diagnostic tests or to mastery tests.

ample, the plans for tests in American history usually call for a few items on social history. Because of the widely prevalent emphasis on political and economic history in American schools, good students as a group know little more about social history than do their less-favored classmates. But even though these items fail to discriminate, they may serve a useful purpose. The stated objectives of most history courses include some mention of social history, so that a test which purports to cover the field adequately can hardly avoid including such items regardless of their effects upon the test statistics. Further, such items emphasize the importance of social history to students and teachers. In the areas where teaching practices are somewhat behind the recommendations of leaders in the field, tests can be of some help in leading the way toward improvement.

The Forms of Test Items

The *form* of an objective test item is determined by the arrangement of words, phrases, sentences, or symbols composing it, by the directions to the examinee for response to it, and by the provision made for recording the response.

A wide variety of item forms have been suggested and used. One manual on test construction lists thirteen major forms with many variations of each. It will not be necessary to consider more than a few important types in this chapter. A few popular forms account for the bulk of all items written. Further, most of the problems that arise in using the more common forms also arise in using other forms, and the principles leading to successful use of the forms discussed here apply also to other forms.

It is worth noting that all objective item forms may be divided into two main classes. On the one hand are items to which the pupil must respond by *supplying* the words, numbers, or other symbols which constitute the response. On the other hand are items to which the pupil responds by *selecting* a response from among those presented in the item. Between the *supply type* and the *selection type* of items there are real differences, but these differences have frequently been misinterpreted so that various forms have been credited with merits and faults that they do not possess.

The forms that will be described here include the short-answer, which represents the supply type, and the true-false, multiple-choice, and matching, which represent the selection type. The item writer needs to be thoroughly familiar with each of these forms. But it is important to note that differences in form do not constitute the only or even the most significant differences among test items. It is now recognized that

many earlier studies of the relative merits of different forms lack significance because they failed to consider characteristics and to control variables that are more important than form.

The Short-Answer Form

The short-answer form is characterized by the presence of a blank on which the examinee writes the kind of answer called for by the directions. This form is no longer widely used aside from informal classroom testing, and even there it has tended to lose favor. It is discussed here because misconceptions concerning its value and applicability still persist, and because it provides an opportunity to emphasize by contrast some of the advantages of other forms. Three varieties of the short-answer form may be illustrated.

1. The question variety

1. Who invented the cotton gin? ____________________
2. How many calories will be required to change eight grams of ice at 0° C. to steam at 100° C.? ____________________

2. The completion variety

1. *Snowbound* was written by ____________________.
2. The body of an insect is divided into three parts: ______________, ______________, ______________. Insects have ______________ antennae and ______________ legs. They breathe by means of ______________.

3. The identification or association variety

1. After each city write the state in which it is located.

Detroit ______________	New Orleans ______________
Chicago ______________	San Antonio ______________
Seattle ______________	Atlanta ______________

The True-False Form

The true-false form consists of a statement to be judged true or false. It is essentially a two-response item in which only one of the possible alternatives is explicitly stated. Several other forms are closely enough related to the true-false form to justify considering them as simple varieties of the true-false form. These are the question which is to be answered yes or no, and the statement which is to be judged correct or incorrect. Illustrations of this form are given below.

1. The true-false variety

This variety consists of a declarative statement that is true or false.

The pressure in a fixed quantity of a gas varies directly as its volume if the temperature remains constant. T F

2. The right-wrong variety

This variety consists of a sentence, equation, or other expression that is to be marked right or wrong depending on whether it is correctly or incorrectly written.

> Glancing down the famous street, signs of every kind were visible. (Sentence structure) R W

3. The yes-no variety

This variety consists of a direct question that is to be answered yes or no.

> Does a ripsaw have larger teeth than a crosscut saw? Y N

4. The correction variety

In this variety the examinee is directed to make every false statement true by suggesting a substitute for the underlined word. This variety combines selection of responses with the supplying of responses to some items.

> The use of steam revolutionized transportation in the <u>17th</u> century. ____

5. The cluster variety

This variety consists of an incomplete stem with several suggested completions, each of which is to be judged true or false.

> The volume of a mass of gas
> 1. tends to increase as the temperature increases. 1. T F
> 2. tends to increase as the pressure increases. 2. T F
> 3. may be held constant by increasing pressure and decreasing temperature. 3. T F
> 4. may be reduced to zero by increasing the pressure and decreasing the temperature. 4 T F

The Multiple-Choice Form

The multiple-choice form consists of the item *stem* (an introductory question or incomplete statement) and two or more *responses* (the suggested answers to the questions, or completions of the statement). In this discussion the correct response or responses will be called the answer(s), and the incorrect responses, the distracters.

The multiple-choice form is by far the most popular one in current use. It is free from many of the weaknesses inherent in other forms. It is adaptable to a wide variety of item topics. While it has often been used to measure superficial verbal associations and insignificant factual details,

and while many examples of poor multiple-choice items can be found, it can also be used with great skill and effectiveness to measure complex abilities and fundamental understandings.

Many variations of the multiple-choice form have appeared. It is convenient to discuss these variations as if each constituted a distinct variety of the multiple-choice form. Actually, of course, these variations affect different characteristics of the item and may be combined in various ways. A single multiple-choice item may thus possess the characteristics attributed to several distinct varieties here.

1. The best-answer variety

This variety of the multiple-choice form consists of a stem followed by two or more suggested responses that are correct or appropriate in varying degrees. The examinee is directed to select the best (most nearly correct) response.

1. Which one of the following factors contributed most to continued inflation in 1948?
   a. Reduction in income taxes.
   b. Bumper crops.
   c. High rate of consumer-purchasing.
   d. Increased payments to veterans.

2. What is the *basic* purpose of the Marshall Plan?
   a. Military defense of Western Europe.
   b. Re-establishment of business and industry in Western Europe.
   c. Settlement of differences with Russia.
   d. Direct help to the hungry and homeless in Europe.

2. The correct-answer variety

This variety consists of an item stem followed by several responses, one or more of which is absolutely correct while the others are incorrect.

3. Who invented the sewing machine?
   a. Singer
   b. Howe
   c. Whitney
   d. White
   e. Fulton

The difference between the best-answer and the correct-answer variety is more one of topic than of form. The name of the inventor of the sewing machine is recorded in history beyond question or doubt. The causes of inflation or the purpose of the Marshall Plan cannot be stated with any such precision.

3. The multiple-response variety

When the item writer is dealing with questions to which a number of clearly correct answers exist, it is sometimes desirable to include two or more correct answers in the choices offered. When this is done, and the examinee is instructed to mark all correct responses, the item variety is designated as the "multiple-response" form.

4. What factor or factors are principally responsible for the clotting of the blood?
   a. Oxidation of hemoglobin.
   b. Contact of blood with injured tissue.
   c. Presence of unchanged prothrombin.
   d. Contact of blood with a foreign surface.

A multiple-response item must also be a correct-answer item. It can never be a best-answer item.

In practice the multiple-response item is not essentially different from the cluster variety of true-false item. This is particularly true if each response is scored as a separate unit, so that the examinee receives one point for each response he does mark that should be marked, and one point for each reponse he does not mark that should not be marked.

It is, of course, possible to score multiple-response items on an "all-or-none" basis. In this method of scoring, the examinee does not receive credit for an item unless he marks all the correct responses and only those.

4. The incomplete-statement variety

Quite frequently the introductory portion of a multiple-choice item (the item stem) consists of a portion of a statement rather than a direct question.

5. Millions of dollars worth of corn, oats, wheat, and rye are destroyed annually in the United States by
   a. rust
   b. mildews
   c. smuts
   d. molds

5. The negative variety

To handle questions that would normally have several equally good answers, item writers sometimes use a negative approach. The responses include several correct answers together with one that is not correct or

which is definitely weaker than the others. The examinee is then instructed to mark the response that *does not* correctly answer the original question, or that provides the least satisfactory answer.

6. Which of these is *not* true of a virus?
   a. It can live only in plant and animal cells.
   b. It can reproduce itself.
   c. It is composed of very large living cells.
   d. It can cause disease.

6. *The substitution variety*

The multiple-choice form has been utilized by item writers in testing a student's ability to express himself correctly and effectively. Samples of originally well-written prose or poetry are systematically altered to include errors in punctuation, spelling, word usage, and similar conventions. Selected words or phrases in these rewritten passages are underlined and identified by number. Several possible substitutions for each critical phrase are provided. The examinee is directed to select the phrase (original or alternative) that provides the best expression.

7. *Selection*

Surely the forces of education should be fully utilized to acquaint youth with the real nature of the dangers to democracy, <u>for</u> (7) no other place offers <u>as good or better opportunities than</u> (8) the school for a <u>rational</u> (9) consideration of the problems involved.

Items

7. a. , for
   b. . For
   c. —for
   d. no punctuation
8. a. as good or better opportunities than
   b. as good opportunities or better than
   c. as good opportunities as or better than
   d. better opportunities than
9. a. rational
   b. radical
   c. reasonable
   d. realistic

7. *The incomplete-alternatives variety*

In some cases, an item writer may feel that the suggestion of a correct response would make the answer so obvious that the item would function poorly or not at all. He may then resort to incomplete or coded alternatives. For example, the examinee may be asked to think of a one-word response and to indicate that response on the basis of its first letter.

8. Thomas Aquinas was an important figure in the development of which school of philosophy?
   1. a to e
   2. f to j
   3. k to o
   4. p to t
   5. u to z

Since the correct answer to this item is "scholasticism," he should mark response 4.

The use of incomplete responses makes possible the objective measurement of such traits as active vocabulary. In tests for this purpose it is essential to force the examinee to think of the appropriate response himself. The following item illustrates this application.

9. An apple that has a sharp, pungent, but not disagreeably sour or bitter, taste is said to be 4* .
   1. R
   2. S
   3. T
   4. U
   5. V

* The figure indicates the number of letters in the word (in this case "tart"). This restriction serves to rule out many borderline correct responses.

Incomplete responses may also be used in arithmetical problems. The student may be directed to mark a choice on the basis of a certain digit in his answer, such as the third digit from the left. The use of incomplete responses for arithmetic problems prevents a student from using the responses as starting points for reverse, short-cut solutions of the problems.

The incomplete-response variety represents a hybrid between the short-answer and multiple-choice form. It has the advantage of perfectly objective scoring. However, like the short-answer form, it is limited to questions for which unique simple correct answers exist. Further, unless the response categories are sharply delimited, credit may be given for wrong answers that happen to fall in the correct response category.

8. *The combined-response variety*

This variety consists of an item stem followed by several responses, one or more of which may be correct. A second set of code letters indicates various possible combinations of correct responses. The examinee is directed to choose the set of code letters which designates the correct responses and to mark his answer sheet accordingly. The following is an example of the combined-response variety.

1. Our present constitution
   a. was the outgrowth of a previous failure.

b. was drafted in Philadelphia during the summer (May to September) of 1787.
c. was submitted by the Congress to the states for adoption.
d. was adopted by the required number of states and put into effect in 1789.

1. a
2. a, b
3. a, b, c
4. b, c, d,
5. a, b, c, d

This variety represents essentially a different method of scoring the multiple-response variety. It is limited to questions for which definitely correct or incorrect responses exist. There is no reason to believe that combined-response items are easier to score or yield more valid scores than straight multiple-response items.

The Matching Form

The matching form of objective test exercise consists of a list of premises, a list of responses, and directions for matching one of the responses to each of the premises. Names, dates, terms, phrases, statements, portions of a diagram, and many other things are used as premises. A similar variety of things may be used for responses. The distinction between premise and response is purely formal. In the present discussion the premises will be identified as those bearing the item number.

Two chief varieties of the matching form are in common use. One is the simple matching exercise illustrated by the following example.

On the blank before each of the following scientific achievements, place the letter that precedes the name of the scientist responsible for it.

_____ 9. Demonstrated the circulation of the blood	a. Louis Pasteur
_____ 10. Demonstrated the statistical approach to human heredity	b. Gregor Mendel
	c. Francis Galton
_____ 11. Conducted crucial experiments on the mechanism of heredity	d. Robert Koch
	e. William Harvey

In items of this type the basis for matching is almost self-evident. The simple matching exercise is chiefly useful for identification of names, dates, structures, and similar associations.

In the exercise above some of the responses do not match any of the premises. Matching exercises having this characteristic, which is often termed "imperfect matching," are widely used. They do not permit the examinee to determine the correct response to the "last" premise by elimination.

Exercises in which each response matches one and only one premise are termed "perfect matching" exercises. These have a serious limitation in that not all of the premises can function as independent items. In a four-premise exercise, correct response to three of the premises guarantees correct response to the fourth. Perhaps the best type of matching exercise is that in which a response may match one, more than one, or none of the premises.

The other chief variety of the matching form is based on the classification of statements.

Directions: In the following items you are to judge the effects of a particular policy on the distribution of income. In each case assume there are no changes in policy that would counteract the effect of the policy described in the item. For each item *blacken* answer space

a. if the policy described would tend to *reduce* the existing degree of inequality in the distribution of income;
b. if the policy described would tend to *increase* the existing degree of inequality in the distribution of income;
c. if the policy described would have no effect, or an indeterminate effect, on the distribution of income.

33. Increasingly progressive income taxes.
34. Confiscation of rent on unimproved urban land.
35. Introduction of a national sales tax.
36. etc.

This variety is well adapted to item topics dealing with explanations, criticisms, and other higher-level learning products.

THE PICTORIAL FORM

Test items based on pictures and graphical representations have not been widely used. This may be due to the fact that many item writers are not artists themselves and do not have the assistance of capable artists. It may also be due to their failure to recognize the advantages which often accrue from the use of pictures and diagrams. Picture test items have been used extensively in the training programs of the armed forces. The illustrative items shown in Figures 2, 3, and 4 have been taken from a bulletin, *How To Make the Picture Test Item* (4), published by the Department of the Army. This bulletin should be consulted for further suggestions on the use of items in this form.

Pictorial items are of two general types—those in which the picture itself presents a problem in interpretation, and those in which the picture serves as an effective means of communicating ideas. Figure 2

illustrates the first type. Figures 3 and 4 illustrate the second. The pictorial material may consist of a photograph, a drawing, a diagram, a graph, or a map. Photographs or drawings usually picture objects, situations, or operations.

Good pictorial items of the interpretive type can be justified as direct measures of useful skills. Those which use pictures as means of communication can be justified on other grounds. In the first place, a picture often provides a more direct and easily understandable expression

50. What is the reading on the vernier shown in the drawing at the right?
A) 10.03
B) 10.13
C) 10.14
D) 11.04

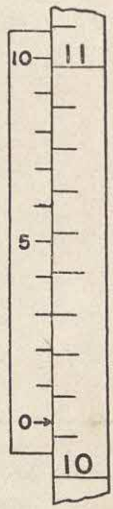
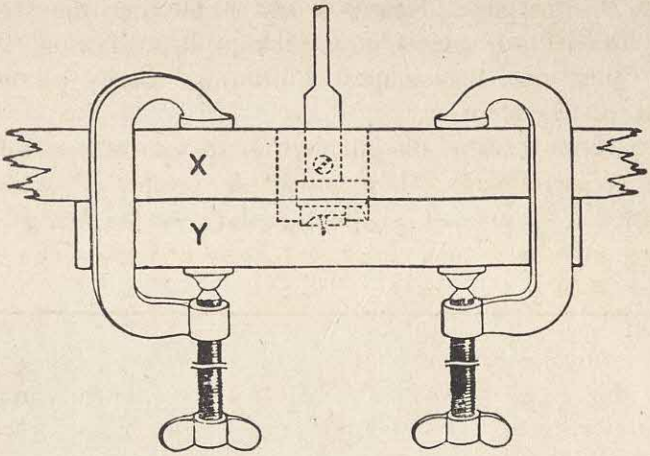


FIG. 2.—Picture items may measure ability to use instruments, read maps, etc. (*From* Department of the Army bulletin [*4*]. Reproduced by permission of the Secretary of the Army.)

than is possible using words alone. In the second, the use of pictures tends to reduce the typically heavy burden of reading comprehension, which often tends to prevent the measurement of nonverbal abilities. The argument that pictures make a test more interesting or attractive has been suggested, but probably merits less weight than the others. While the use of pictures has been generally neglected in test construction, a few enthusiasts have overused them. If a problem can be clearly expressed in a few simple words, it is not ordinarily improved by the inclusion of a picture.

Some words are involved in almost every pictorial item. The suggestions made later for the construction of written items in various forms



21. The board Y is clamped to the board X in order to
    A) prevent board X from splintering when boring
    B) make the boring easier
    C) act as a stop for the bit
    D) make withdrawal of the bit easier

FIG. 3.—Picture test item containing groups of objects to show relationship. (*From* Department of the Army bulletin [*4*]. Reproduced by permission of the Secretary of the Army.)

48. Which is the best method of suspending a splice in spiral four cable?

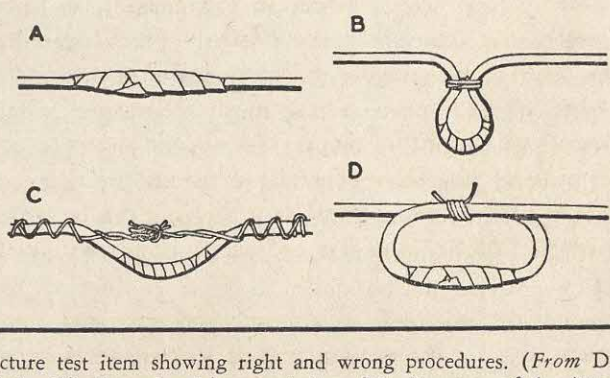


FIG. 4.—Picture test item showing right and wrong procedures. (*From* Department of the Army bulletin [*4*]. Reproduced by permission of the Secretary of the Army.)

will apply in general to items based on pictures. The one additional factor which needs to be considered is the quality and appropriateness of the pictorial material used. Unless skillfully drawn so that the objects or situations represented are easily recognized and the essential elements receive proper emphasis, the drawing may defeat its own purpose.

Characteristics of Objective Test Items

Validity

The validity of an item for a given purpose depends considerably upon the idea it involves and upon the skill with which it is constructed. There is, however, a persistent belief that certain item forms are intrinsically more valid than others. Short-answer items are often credited with high validity because they require the examinee to supply an answer of his own. Those who hold this point of view argue that the ability to supply an answer is more useful and indicates higher achievement than the ability to select one from suggested alternatives. Both of these arguments need examination.

In the first place, the arguments appear strongest when one has in mind an item intended to test memory for specific facts. A short-answer item involving spelling or addition is clearly a more direct and, hence, intrinsically more valid measure of achievement in those areas than a choice-type item is. On the other hand, items in short-answer form have no such apparent advantage over choice-type items for measuring judgment or reasoning ability. In practice it has been found much easier to differentiate good from poor reasoners on the basis of their choice of conclusions than on the basis of their written statements of those conclusions.

In the second place, choice-type items only *appear* to give the examinee too much help. When critics ask how an examinee can fail to get the correct answer when it is printed in plain sight, the experienced test constructor answers that many examinees do actually fail to select the correct response. Item writers can prepare fair and clearly stated choice-type items in most areas that are difficult enough to discriminate between those who have and those who lack a given aspect of achievement. Contrary to naïve expectation, many items become more, instead of less, difficult when transformed from short-answer to multiple-choice form. A generally correct notion that would be quite adequate for response to a short-answer item will not be adequate for response to a choice-type item that

requires the examinee to make precise discriminations among several plausible alternatives. The following item illustrates this point.

What causes the diaphragm of a telephone receiver to vibrate?
1. Changes in voltage caused by a transformer.
2. Changes of current strength in an electromagnet.
3. Changes in position of carbon granules pressing against the diaphragm.
4. Changes in the position of the condenser.

When this question is used alone as a short-answer item, it requires only such simple general answers as "electrical currents," "magnetism," or "an electromagnet." When transformed into multiple-choice form, much more detailed understanding is required.

In the third place, while a person who can on demand produce an answer demonstrates higher achievement *of a particular type* than one who cannot, it by no means follows that ability to produce answers is *generally a more useful type* of achievement than ability to make right choices. In fact, the reverse is probably true. Success in many occupations and activities is more closely related to the ability of making a wise choice among known alternatives than of producing new ideas.

Finally, whatever case can be made for short-answer items on other grounds is weakened seriously by the ambiguity inherent in them, and by the consequent difficulty of scoring them fairly. These points are discussed on page 207.

Errors Due to Guessing

Many users of objective achievement tests suppose that test scores derived from choice-type items are subject to serious errors due to guessing. Some also believe that the chief weakness of the true-false item is the frequency with which poorly prepared examinees make high scores by "coin tossing." It can hardly be denied that some chance response to choice-type items is possible, but the seriousness of the problem has been exaggerated. In this section only those aspects of guessing which are related to the forms of objective test items will be considered. The chapter on test administration and test scoring (pages 365–69) will discuss the matter in greater detail.

In considering this problem, it is helpful to distinguish three bases for response to test items: certain knowledge or misinformation, informed guesses, and blind chance. Responses based on certain knowledge introduce no error to test scores. Responses based on blind chance contribute only error. The amount of error introduced by guesses depends upon the

amount of information on which they were based. Guesses based on almost no information are practically equivalent to chance responses. Those based on almost complete information may be practically the same as certain responses. It is a tenable hypothesis that all but the wildest guesses contribute more to the true variance of test scores than to their error variance, and hence constitute useful responses from the standpoint of measurement. Under this hypothesis, the only harmful "guesses" are those based on almost no information, which are not really guesses at all but chance responses.

The amount of error introduced by chance responses depends in part upon the frequency of such responses. The conditions under which response by chance is attractive to the typical examinee need not occur commonly. An examinee who is interested in making the highest possible score and who has sufficient time seldom prefers a blind response to one based on consideration of all available information. Both motivation and time limits are at least partially under the control of the test administrator. If they are handled properly, the frequency of chance response and, hence, the amount of error introduced by those responses can be made quite small.

While the probability of a correct response to a single true-false item by chance is twice that of the corresponding probability for a four-choice item, it by no means follows that scores on a true-false test involve twice as much error as those from four-choice items. In fact, if chance responses are given to the same number of true-false and four-choice items, the error variance due to chance response is only one-third larger in the case of the true-false items than in the case of the four-choice items. Further, the effect of this error variance due to chance on the reliability of the test scores depends on the total variance of the obtained scores and on the potency of other sources of error in the scores. If empirical evidence of these factors is lacking, it is hardly justifiable to assume that true-false items are inevitably subject to much greater error through chance response than are four-choice items.

Many test users do not recognize that the possibility of obtaining a high relative score on a well-constructed true-false test purely by chance is very small. Suppose (quite unrealistically) that 100 examinees respond blindly to each item of a 200-item true-false test, and that each examinee's score consists of the number of items he answers correctly. The expected mean of these chance scores would be 100. Assuming the distribution follows the binomial law, none of the examinees would be *expected* to

receive scores of 120 or higher. In the long run chance scores larger than 119 would occur less than three-tenths of 1 percent of the time. By giving due attention to the difficulty of the items, the test constructor can produce a test yielding a distribution of scores whose lower limit would be above 120. Under these conditions the frequency with which an examinee could make a "good" score by "coin tossing" would be too small to concern the test constructor.

Applicability of Various Forms

Short-answer

Experience has shown that it is nearly impossible to phrase short-answer items on certain essential topics so that the same correct responses will be made by all those who know the answers. Item writers are often surprised by the variety of correct responses which appear to questions for which they had conceived only one answer. Most words and phrases have many synonyms and near-equivalents. Also, it is often possible to fill the blank of a short-answer item with words which are appropriate, but which are remote from the intent of the item writer. In the item

> Columbus discovered America in ________.

the examinee may appropriately respond with either "1492" or "the *Santa Maria*." Even mathematical problems which should yield precisely determinable answers cause trouble unless explicit directions have been given concerning the form and precision required in the answer.

As creditable responses to a test item multiply, the scoring key becomes increasingly cumbersome, the scoring procedure more time-consuming, and the obtained scores less reliable. Further, it becomes inadvisable to entrust the scoring to clerks who lack mastery of the subject matter involved. Practically speaking, only items involving simple computations or simple verbal associations can be handled with satisfaction in the short-answer form. These difficulties, of course, are less serious in classroom testing where the tests are scored by a competent teacher and where speed of scoring may be a secondary consideration.

True-false items

True-false items should be based only on statements that are absolutely and unambiguously true or false. A relatively small proportion of the significant statements that can be made on any subject satisfy this criterion. To meet the standard of absolute truth, a statement must be so precise

in phrasing and so universal in application that it requires no additional qualifications and admits of no possible exceptions.

If statements that are only approximately true are presented as items in a true-false test, they pose a difficult and unreasonable problem to the examinee. Not only must he know *to what extent the statement is true,* but he must also guess what *degree of untruth* will be tolerated by the scorer.

Consider the problem of response to these hypothetical true-false statements.

1. The numerical value of pi is 3.
2. The numerical value of pi is 3.1416.
3. The numerical value of pi, correct to four decimal places, is 3.1416.

If presented in separate tests, it is safe to guess that most informed examinees would mark item 1 *false* and item 2 *true.* Yet both items are basically alike in being approximations. Each is true as far as it goes. Only by further qualification, as in item 3, can the item be made absolutely true.

The difficulty involved in securing absolutely true statements for use in true-false items may be illustrated by this example.

4. Calcium chloride attracts a film of moisture to its surface and gradually goes into solution.

This item is true if and only if solid calcium chloride is in an atmosphere containing moisture.

The following item is also questionable.

5. No satisfactory explanation has ever been given for the migration of birds.

Many explanations have been offered for the migration of birds. It is conceivable that one of them will ultimately prove correct. Further, the item does not specify who must be "satisfied" by the explanation.

Ambiguity is also present in this statement:

6. The nourishment assimilated by the body depends upon the amount of food eaten.

No one would argue that there is a *perfect* relationship between assimilation and intake over all possible values of intake. It would be equally absurd to claim that there is *no* relationship.

Although changes in wording would improve some of the foregoing items, the basic difficulty is not one of clear expression. Nor can the

difficulty be avoided (although it may be reduced somewhat with some items) by the use of qualified response categories such as, "completely true, mostly true, mostly false, completely false." Instead, the ambiguity is an inevitable result of the attempt to apply an abstract standard of absolute truth to statements whose truth is relative, conditional, or approximate.

This requirement of absolute correctness tends to limit the applicability and the validity of items in true-false form. Many important outcomes of instruction are generalizations, explanations, predictions, evaluations, inferences, and characterizations. Since these things often cannot be expressed in statements which are precisely and universally true, they cannot be tested effectively by true-false items. Analyses of responses reveal that it often is the brighter, better-informed students who sense the need for qualifications and the possibility of exceptions to statements presented as true-false items. When this happens, the item loses validity, and may even have negative validity.

Recognition of this limitation has led some item-writers to use the true-false form only to test memory for simple facts. This is a needless restriction. With certain topics, well-constructed true-false items can stimulate fairly complex reasoning processes. Consider, for example, these statements:

1. If a square and an equilateral triangle are inscribed in the same circle, the side of the square is longer than the side of the triangle.
2. It is possible for an erect man to see his entire image in a vertical plain mirror one half as tall as he is.

An ingenious item writer can borrow or invent many similar situations. When correctly understood, they lead to an unequivocal response of "true" or "false." If the examinee has not encountered the same situation before, his correct response indicates ability to handle a complex reasoning process.

The multiple-choice form

The multiple-choice form is widely applicable. An item presented in short-answer, true-false, or matching form can always be converted into one or more multiple-choice items. Frequently this conversion improves the effectiveness of the item. Since there is only one correct response which an examinee can make to a multiple-choice item, the difficulty and subjectivity of scoring which plague the short-answer form are avoided. Since the multiple-choice form is adapted to the best-answer approach, it

avoids the ambiguity associated with the application of a standard of absolute truth, which constitutes the chief weakness of the true-false form. Since each multiple-choice item may be independent, the problem of finding a number of parallel relationships, which frequently causes difficulty in the matching form, may be avoided.

The matching form

The matching exercise is poorly adapted to unique topics or test situations. Since each of the responses should have some plausible relationship to each of the premises, both responses and premises must be relatively homogeneous. But many significant topics are unique and cannot be conveniently grouped in homogeneous matching clusters. Consider, for example, the difficulty of incorporating items like the following, all of which deal with prices, in any homogeneous set of premises and responses.

1. Why did the prices of radios and tires decline during 1948 at a time when the prices of other consumer goods were still rising?
   a. Costs of manufacture declined.
   b. There is normally more competition in these industries.
   c. There was organized consumer resistance to high prices of these commodities.
   d. The backlog of consumer demand disappeared.

2. Buyers' strikes against higher prices have demonstrated
   a. their effectiveness in the case of foods but not of other products.
   b. their effectiveness in rural areas but not in cities.
   c. their relative ineffectiveness in reducing prices.
   d. their effectiveness in reducing wholesale but not retail prices.

3. Which one of the following factors contributed most to continued inflation in 1948?
   a. Reduction in income taxes.
   b. Bumper crops.
   c. High rate of consumer purchasing.
   d. Increased payments to veterans.

It may also be pointed out that use of the matching form may exert an undesirable influence on the distribution of emphasis in the test. An item writer may have in mind a particular date-event relationship that is of considerable importance. It occurs to him that there are similar date-event relationships that, though not of similar importance, might be included in a single matching exercise. As a result, the test finally produced may have excessive emphasis in this area, while other important

aspects of achievement, which do not lend themselves well to grouping in a matching test, may be neglected.

Ease of Construction

The difficulty of constructing an objective test item depends far more on the level of quality demanded than on the form in which the item is cast. When faulty items are accepted uncritically, the construction of items in any form appears quite easy. Further, the difficulty of constructing an item depends on the topic with which it deals. For example, items measuring memory for facts are far easier to construct than items measuring understanding.

Popular belief, borne out by casual experience, is that short-answer and true-false items are easier to construct than multiple-choice items. Short-answer items resemble the informal questions with which the teacher is familiar. True-false items resemble the declarative statements which abound in textbooks. It thus appears quite a simple matter to build objective short-answer and true-false items on the basis of these handy materials. Actually the borrowing of such questions and statements is likely to yield unscorable or ambiguous items.

The chief problem in the construction of short-answer items is to find questions each of which has a single correct answer. The search for such questions too often leads the item writer to concentrate on verbal associations and factual details and to avoid interpretations and other complex relationships. The practice of producing one variety of short-answer item by removing words from selected statements often yields vague, multiple-answer items. As an extreme example, consider this statement and the short-answer items derived from it.

1. Tobacco is grown in the South.
2. ________ is grown in the South.
3. Tobacco is grown in the ________.

While the original statement was perfectly correct, it is obvious that the items derived from it can be completed correctly in many different ways.

The chief problem in the construction of true-false items is to limit the ideas to those that can be expressed as absolutely true or false statements. A closely related problem is to word the statement so that it is unambiguous without being obviously true or false to the uninformed.

The chief problem in the construction of multiple-choice exercises is to express a question or problem clearly in the item stem, to phrase the correct response defensibly, and to find attractive distracters which

will permit the item to discriminate between those who have and those who lack the achievement involved.

The chief problem in the construction of matching exercises is to find homogeneous premises and homogeneous responses for which a meaningful basis for matching exists.

Ease of Scoring

One of the chief reasons for the wide acceptance of objective test forms is the ease with which they can be scored. Choice-type items are much easier to score than short-answer items. Responses to true-false, multiple-choice, or matching items can be indicated by marking various positions on a separate answer sheet. These answer sheets can then be scored rapidly by clerical inspection if stencil keys are used or by electrical machines if responses have been marked with special pencils.

The efficiency, and consequent widespread use, of the electrical test-scoring machine, or of stencil keys for clerical scoring, has tended to favor certain item forms at the expense of others. Multiple-choice, true-false, and to a lesser extent matching exercises are well adapted to this mechanical type of scoring. Some writers have viewed this development with alarm, fearing that useful old forms may be abandoned and useful new item forms may not be developed because they do not fit mechanical-scoring requirements. Whether or not these fears are justified is difficult to say. One may take the position that more is likely to be gained in educational measurement through improved application of existing item forms than through the development of new forms, and that, except in rare instances, the unadaptability of certain item forms to mechanical scoring represents no real loss to educational measurement. On the other hand, one should certainly not ignore the possibilities of any new evaluation technique solely because it does not appear suitable for mechanical scoring.

The use of mechanical devices as aids in test scoring has been a cause of both amazement and critical comment by individuals not directly concerned with measurement. They have asserted that a mechanical device can be useful only in evaluating mechanical performances. If the test is poorly constructed this assertion may be true—but no more so than if other scoring methods were used. Competent writers can produce items which require exceedingly complex and original thought processes whose outcome can be recorded in a simple mechanical fashion. The simplicity of the record of the outcome is *not* a necessary indication of simple procedure in arriving at the answer.

Suggestions for Item Writing

General Suggestions

1. *Express the item as clearly as possible.*

The production of good test items is one of the most exacting tasks in the field of creative writing. Few other words are read with such critical attention to implied and expressed meaning as those used in test items. The problem of ambiguity in objective test items is particularly acute because each item is usually an isolated unit. Unlike ordinary reading material, in which extensive context helps to clarify the meaning of any particular phrase, the objective test item must be explicitly clear in and of itself. The power of an item to discriminate between the competent and incompetent may be seriously limited by lack of clarity. Except in the case of certain types of intelligence or reading test items, the difficulty of an item should arise from the problem involved rather than from the words in which it is expressed. Test items should not be verbal puzzles. They should indicate whether the student can produce the answer, not whether he can understand the question.

Lack of clarity in a test item may arise from inappropriate choice and awkward arrangement of words. It may also arise from lack of clarity in the thinking of the person who wrote it. Many ideas for test items are vague and general at first. Before emerging in final form, they need critical examination and revision. In this process, clarification of ideas goes hand in hand with improvements in wording.

It is difficult to provide a list of specific rules which, if followed, will guarantee clarity. The things that must be done and the things that must be avoided are numerous and varied. Further, their application in specific situations is a matter calling for expert judgment. It is worth noting, however, that many of the suggestions made in the remainder of this section may be considered as elaborations of this first and most important suggestion.

2. *Choose words that have precise meaning wherever possible.*

Lack of clarity in an item frequently arises from inappropriate word choices. Many commonly used words and phrases have no precise meaning. Others have no meaning that applies accurately in the context in which they appear.

In the following item, the words "yielded," "unofficially," and "points," are vague.

1. In the 1948 presidential campaign, Truman yielded unofficially to Dewey on which of these points?

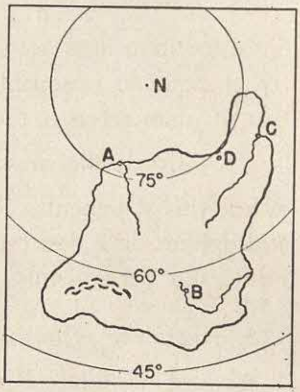
a. That the un-American activities investigations were not direct attacks on the Democratic party.
b. That the tone of the political campaign should be kept on a high level.
c. That the civil rights program should be abandoned.
d. That there should be an equal balance of Democrats and Republicans in Congress to insure sound legislation.

3. *Avoid complex or awkward word arrangements.*

The following item, dealing with the interpretation of a map of the Northern Hemisphere, was very awkwardly worded on the first attempt:

1. What is the relative length of the shortest path between cities *A* and *C*, and the North Pole?
   a. They are approximately equal.
   b. That from *A* is slightly longer than that from *C*.
   c. That from *A* is slightly shorter than that from *C*.
   d. That from *A* is twice as long as that from *C*.

When revised, with the addition of two cities among the distracters, the item reads as follows.



2. Which city is closest to the North Pole?
   a. City A
   b. City B
   c. City C
   d. City D

While this may not be the place for a discourse on rhetoric, a few suggestions may be made to illustrate the general point involved. The structure of sentences used should be as simple as possible. It is often advantageous to break up a complex sentence into two or more separate sentences. A qualifying phrase should be placed near the term it qualifies. In general, it is desirable to make clear the point of the question early in the sentence and add qualifications or explanations later. Finally, it is often helpful for the item writer to ask himself, "Just what is the point of this item?" In the answer to this question an item writer may find a simpler, more direct wording for his item. Or he may discover that it has little point, or one that is not worth testing.

4. *Include all qualifications needed to provide a reasonable basis for response selection.*

Frequently an item writer does not state explicitly the qualifications that exist implicitly in his own thinking about a topic. He forgets that a differ-

ent individual, at another time, needs to have these qualifications specifically stated. The following item illustrates this point.

1. If a ship is wrecked in very deep water, how far will it sink?
   a. Just under the surface.
   b. To the bottom.
   c. Until the pressure is equal to its weight.
   d. To a depth which depends in part upon the amount of air it contains.

A number of capable students selected response *d,* instead of the intended correct response *b,* because they considered the possibility (which the writer failed to exclude) that a wrecked ship might not sink completely but remain partly submerged. In that case, response *d,* while not good, is the best response available.

2. The greatest loss in hail storms results from damage to
   a. livestock
   b. skylights
   c. growing crops

Since the item does not specify loss *to whom* or in *what part* of the country, there is no reasonable basis for response unless the examinee assumes that the item refers to the United States as a whole.

3. Calcium chloride attracts moisture and goes into solution.

While this statement is true in the typical atmosphere, it is not true in the atmosphere of a desiccator. Capable examinees will recognize this possibility, and perhaps miss the item because of their superior ability.

5. *Avoid the inclusion of nonfunctional words in the item.*

A word or phrase is nonfunctional when it does not contribute to the basis for choice of a response. Unnecessary words and phrases frequently fit easily in the wording of an item but prove on examination to be completely irrelevant to the decisions that must be made in selecting the answer.

Sometimes an item writer may include an introductory statement in an effort to strengthen the apparent appropriateness or significance of the item. The following illustration, taken from a test of contemporary affairs, includes such a statement.

1. While many in the U.S. fear the inflationary effects of a general tax reduction, there was widespread support for a Federal community-property tax law under which
   a. husbands and wives could split their combined income and file separate returns.

b. homesteads would be exempt from local real estate taxes.
c. state income taxes might be deducted from Federal returns.
d. farmland taxes would be lower.

In order to answer this item, it is necessary to know only in what way the community-property laws affect tax computation. This question can be brought into sharper focus by rewording the item as follows, eliminating reference to "fear of inflation," "general tax reduction," or "widespread support."

2. Community-property tax laws permit
   a. husbands and wives to split their combined income and file separate returns.
   b. homesteads to be exempt from local real estate taxes.
   c. state income taxes to be deducted on Federal returns.
   d. farmland taxes to be lowered.

Some introductory statements that are not strictly necessary may occasionally be justified as "window dressing" that helps to clarify the point of an item or to establish its importance. This use is illustrated in the following item.

3. The pollution of streams in the more populous regions of the United States is causing considerable concern. What is the effect, if any, of sewage on the fish life of a stream?
   a. It destroys fish by robbing them of oxygen.
   b. It poisons fish by the germs it carries.
   c. It fosters development of non-edible game fish that destroy edible fish.
   d. Sewage itself has no harmful effect on fish life.

If, however, the irrelevant material is put in to make the answer less obvious or to mislead the examinee into choosing an otherwise weak distracter, the result is totally bad. Not only does it tend to destroy the validity of the item as a measure of what it is intended to measure, but it weights undesirably the factor of reading comprehension as a determiner of correct response. In general, items should be kept as short as is consistent with clear statement. Examinees vary widely in speed of reading, and the inevitable advantage given to the rapid readers in most objective tests should not be unnecessarily increased.

6. *Avoid unessential specificity in the stem or the responses.*

General knowledge is knowledge that may be applied in a variety of specific situations. The superior value of general knowledge over specific knowledge has long been recognized and should be reflected in tests wher-

ever possible. The following item is undesirably specific both in the problem presented and in the alternative responses.

1. What percent of the milk supply in municipalities of over 1,000 was safeguarded by tuberculin testing, abortion testing, and pasteurization?
   a. 11.1%
   b. 20.3%
   c. 31.5%
   d. 51.9%
   e. 83.5%

A correct response to this item simply indicates that the student remembers what he has read or heard in class. It does not indicate that he has a general conception of the present status of safeguards to the purity of milk. Items stated in this way are powerful incentives to rote-learning and may unfairly penalize a student whose study habits are basically sound and effective.

A better approach is employed in the following item:

2. What techniques have been widely adopted in an effort to safeguard the purity of city milk supplies?
   a. Only pasteurization of milk.
   b. Only the elimination of diseased cows.
   c. Neither of the preceding has been widely adopted.
   d. Both have been widely adopted.

It should not be assumed on the basis of the foregoing discussion that the more general an item is, the better it is. General statements are ordinarily less precise than specific statements. They require more qualifications and admit more exceptions. It is less easy to state an acceptable answer to a general question. Thus, there are disadvantages as well as advantages in the use of general questions. The item writer must strike a careful balance between the two in order to produce the best possible test item.

7. *Avoid irrelevant inaccuracies in any part of the item.*

Irrelevant inaccuracies are usually unintentional. Even though they may have nothing to do with the selection of the correct response or the elimination of incorrect responses, their inclusion is undesirable for two reasons. In the first place, they reflect unfavorably on the information and ability of the item writer. In the second, they may become fixed in the mind of the student as true statements, suggesting or re-enforcing erroneous ideas. Consider the following example.

1. Why did Germany want war in 1914?
   a. She was following an imperialistic policy.
   b. She had a long-standing grudge against Serbia.
   c. She wanted to try out new weapons.
   d. France and Russia hemmed her in.

In many respects this is a significant and well-constructed item. It does, however, suggest something that has not been established historically. It is unlikely that Germany was actively seeking war in 1914. A more reasonable interpretation is that she was seeking certain goals and would accept war rather than give up those goals. A somewhat similar fault is illustrated by the following item.

2. Studies of general intelligence indicate young people should be first permitted to vote at what age?
   a. 16 c. 20
   b. 18 d. 22

This item implies that general intelligence is the only factor to be considered in determining minimum voting age. Few political scientists or sociologists would accept this view. The real basis for response in this item is the examinee's information that mental ability, as defined in intelligence testing, does not increase much in the typical individual after the age of eighteen. But it is undesirable to imply that this factor alone should determine the minimum age for voting.

8. *Adapt the level of difficulty of the item to the group and purpose for which it is intended.*

Recommendations concerning item difficulty and the statistical basis on which they rest will be discussed in another chapter. It is sufficient here to point out that the usefulness of a test item depends in no small measure upon the appropriateness of its difficulty for the group of examinees who will take it.

While subjective judgments of item difficulty have been found not highly accurate, an individual who is well acquainted with the general level of ability of the examinees and their typical performance on similar items can do a useful job of judging item difficulties.

In connection with the adjustment of item difficulty, two pitfalls need to be avoided. The first is the application of a "minimum essentials" concept to an achievement test. According to this concept, a test should include only those questions that *all* students should, according to the objectives of instruction, be able to answer correctly. A test composed of items meeting this requirement will almost certainly be too easy to discriminate

clearly between different levels of ability. The second mistake is the selection of items from the standpoint of what the ideal student should know rather than in terms of what the typical student does know. An item writer who is unrealistic about the extent to which typical students appreciate the fine points of a subject is likely to produce unreasonably difficult items.

The difficulty of an objective test item is not determined solely by the idea on which it is based. A writer can often construct several items on the same general idea that differ widely in difficulty. Actually, of course, the items may be testing different abilities, but it is difficult if not impossible to define the difference in abilities measured in terms other than the specific differences between the items themselves.

Thus, it is a mistake to interpret item analysis data as indicating, for example, that a certain percent of students in a specified group understand osmosis, or the formation of hailstones. All one is justified in saying is that a certain percent are able to answer correctly a particular item on osmosis, or a particular item on the formation of hailstones. Low scores on a test indicate either low achievement or difficult items or both. High scores may reflect easy items rather than high achievement.

An item writer can control partly the difficulty of multiple-choice and matching items by adjusting the homogeneity of responses or by making use of compound responses. The more homogeneous the responses to a multiple-choice item, the greater the difficulty of the item. Where responses are sufficiently heterogeneous, selection of the best is reasonably easy. Two versions of the same vocabulary item will illustrate this point.

1. His *gaunt* companion
   a. beautiful
   b. healthy
   c. haggard
   d. youthful

2. His *gaunt* companion
   a. ugly
   b. ill
   c. haggard
   d. aged

Obviously, the responses in the second item are much more nearly alike, so that a more specific and hence, presumably, less widely possessed knowledge is required to select the best. Item writers who have a tendency to look for fine distinctions and thus tend to produce relatively difficult items may frequently "ease up" the items by substituting distracters that

are less like the answer. On the other hand, the item writer should be aware that such changes may also change the thing which the item is measuring. The hypothetical universe of potential items from which the test constructor draws his sample is usually large and representative enough so that it is not necessary to include many items which are inappropriate in difficulty for the group being tested. But if the alteration or elimination of certain items systematically reduces the emphasis on important objectives, the changes are definitely undesirable.

Compound responses may be used to simplify an item which would otherwise be too difficult. This possibility is illustrated in the following items.

3. The Nobel Peace Prize was awarded in 1947 to a
   a. group of scientists who refused to work on the atomic bomb.
   b. committee of church members organized to help war victims everywhere.
   c. group of newspapermen who strongly supported the United Nations.
   d. group of lawyers who prosecuted Nazi criminals.

4. When the precipitate of ferrous hydroxide is exposed on a filter paper what happens to it? Why?
   a. It turns red due to oxidation.
   b. It turns green due to reduction.
   c. It turns black due to decomposition.
   d. It remains unchanged due to its stability.

A response may be selected to item 3 by one who knows either which group received the prize or the basis on which it was awarded. Likewise, in item 4, a response may be selected on the basis of knowledge of the outcome of the experiment or on the basis of understanding of the fundamental principles involved.

Compound responses may also be used to increase the difficulty of an item by requiring the examinee to demonstrate two abilities in choosing a response. One common illustration of this possibility is found in items that require both an answer and an explanation.

5. If a tree is growing in a climate where rainfall is heavy, are large leaves an advantage or a disadvantage?
   a. An advantage, because the area for photosynthesis and transpiration is increased.
   b. An advantage, because large leaves protect the tree during heavy rainfall.
   c. A disadvantage, because large leaves give too much shade.
   d. A disadvantage, because large leaves absorb too much moisture from the air.

9. *Avoid irrelevant clues to the correct response.*

The effect of irrelevant clues may be to make the item easier as a whole or, which is more serious, may change the basis upon which the item discriminates. If all students notice the clue and all respond correctly on the basis of it, the item becomes nondiscriminating and, hence, useless. If only the more capable examinees utilize the clue and all others overlook it (that is, if ability to use the clue is highly related to the ability the test is intended to measure), the item may not be seriously weakened. More commonly, however, a number of examinees who would not normally be able to choose the correct response notice the clue and respond correctly on the basis of it. In this case the presence of the clue definitely weakens the item.

Irrelevant clues may be of several varieties. Clues may sometimes be provided by pat verbal associations.

1. What does an enclosed fluid exert on the walls of its container?
   a. Energy c. Pressure
   b. Friction d. Work

It is necessary only to know that "exert" is commonly used with "pressure" to answer the question correctly. Another type of irrelevant clue is provided by grammatical construction.

2. Among the causes of the Civil War were
   a. Southern jealousy of Northern prosperity.
   b. Southern anger at interference with the foreign slave trade.
   c. Northern opposition to bringing in California as a slave state.
   d. differing views on the tariff and constitution.

Quite obviously the item stem calls for a plural response, which occurs only in *d*.

An item writer may provide irrelevant clues to the correct responses by consistently stating them more precisely and at greater length than the foils.

3. Why were the Republicans ready to go to war with England in 1812?
   a. They wished to honor our alliance with France.
   b. They wanted additional territory for agricultural expansion and felt that such a war might afford a good opportunity to annex Canada.
   c. They were opposed to Washington's policy of neutrality.
   d. They represented commercial interests which favored war.

It should be noted that a single item cannot adequately illustrate this defect. If other items include *incorrect* responses phrased as elaborately and precisely as this correct response, the effect of the clue will be offset.

However, there is a natural tendency for an item writer to be more careful in phrasing the correct response than in phrasing the distracters. Precision of phrasing has frequently been used by alert but poorly informed examinees as the basis for choosing the correct response.

Irrelevant clues may also be provided by *any systematic formal differences* between the answer and the distracters. Some item writers tend to place the answer in one favored position among the several alternatives. If, for example, an examinee observes that the third response given is most frequently the answer, or that the first response is seldom correct, he may use such observations as a basis for successful guesses on other items.

One of the most frequently encountered types of irrelevant clue is provided by common elements in the item stem and in the answer. An obvious example of this is provided in the following item.

4. What led to the formation of the States Rights Party?
   a. The level of Federal taxation.
   b. The demand of states for the right to make their own laws.
   c. The industrialization of the South.
   d. The corruption of many city governments.

Common elements in stem and a response are frequently far less obvious than this, but they may spoil the effectiveness of an otherwise well-constructed multiple-choice item.

Occasionally an item writer will provide clues by inadvertently including interrelated items, so that the statement of one question, or its responses, provides a direct clue to the answer of another. For example:

5. The term "biological warfare" refers to
   a. the struggle of living things to survive.
   b. the use of disease-producing organisms to defeat or weaken an enemy.
   c. the conflict between evolutionists and anti-evolutionists.
   d. the use of drugs to help save lives in combat.

6. How far has the use of disease-producing organisms in warfare been developed?
   a. The idea has been suggested but not developed.
   b. "Biological warfare" has been developed somewhat but it is not yet ready for use.
   c. Techniques of "biological warfare" are developed and ready for use.
   d. "Biological warfare" was used extensively in World War II, especially by saboteurs.

The stem and responses of item 6 provide a direct suggestion of the answer to item 5.

Interlocking questions are undesirable because they cannot be depended upon to measure what they were intended to measure. The best safeguard

against errors of this type is careful rereading of a test as a whole, with particular attention to the relationships among the items.

Words like "all," "none," "certainly," "never," and "always" have been designated "specific determiners." They tend to operate as irrelevant clues to the correct response, especially when used in true-false items. Statements including them are predominately false. These words may be noted by clever examinees and used as a basis for choosing the correct response even when the fundamental knowledge or ability involved is lacking. Thus, specific determiners change the basis upon which an item discriminates and often destroy its usefulness. The falsity of the following sample true-false items is obvious because of the specific determiners included.

1. *All* diseases require medicine for their cure.
2. If water is brought to a boil, *all* bacteria in it will *certainly* be killed.
3. If drinking water is clear, it is *always* safe for drinking.

Irrelevant clues provided by pat verbal associations or by common elements in the item stem and one of the responses may be used constructively in multiple-choice items. Deliberate planting of clues of this type in the foils tends to defeat the rote-learner or to make the distracters highly attractive to those whose knowledge is superficial. This practice increases the power of the item to discriminate between real and simulated achievement.

10. *In order to defeat the rote-learner, avoid stereotyped phraseology in the stem or the correct response.*

A rote response is here defined as one in which words are used with no clear conception of their meanings. Rote responses are usually based on verbal stereotypes. The following questions (with intended answers in parentheses) illustrate the opportunities that certain items provide for response by rote.

What is the biological theory of recapitulation?
(Ontogeny repeats phylogeny.)

Who was the chief spokesman for the *"American System"?*
(Henry Clay)

What were the *staple crops* in the colonial South?
(tobacco, rice, indigo)

In the first item the verbal stereotype is in the phrase "Ontogeny repeats phylogeny" and its association with the term "recapitulation"; in the second, "Henry Clay's American System"; in the third, "staple crops." To

demonstrate the stereotyped character of these phrases, the reader is invited to attempt the clarification of the phrases "Ontogeny repeats phylogeny," the "American System," and "staple crop." It is true that some examinees may know the meanings of these phrases, but it is also obvious that the items do not *require* this knowledge for successful response.

The emphasis supply-type (see page 193) items place on some *unique* word or phrase as *the* answer makes them particularly subject to response by rote. When this form is used, it is almost impossible to avoid giving credit to the "lesson learner" who knows the words, whether or not he knows the meaning. Any correct answer supplied, however obviously stereotyped, must be given full credit. With choice-type items, on the other hand, it is possible to avoid such verbal stereotypes among the correct responses, or to work them into the distracters. The extent of rote-learning and the harm which it does were discussed by Lindquist at some length in *The Construction and Use of Achievement Examinations* (*3*), from which the following quotation is taken.

> Consider the following illustrations. The items given below were included in a battery of tests administered to a random sample of 325 physics students in Iowa high schools.
>
> 1. What is the heat of fusion of ice in calories?
> (Answered correctly by 75 per cent of the pupils.)
>
> 2. How much heat is needed to melt one gram of ice at 0° C.?
> (Answered correctly by 70 per cent of the pupils.)
>
> 3. Write a definition of heat of fusion.
> (Answered correctly by 50 per cent of the pupils.)
>
> 4. The water in a certain container would give off 800 calories of heat in cooling to 0° C. If 800 grams of ice are placed in the water, the heat from the water will melt
> (1) all the ice
> (2) about 10 grams of the ice
> (3) nearly all the ice
> (4) between 1 and 2 grams of the ice
> (Answered correctly by 35 per cent of the pupils.)
>
> 5. In which of the following situations has the number of calories exactly equal to the heat of fusion of the substance in question been applied?
> (1) Ice at 0° C. is changed to water at 10° C.
> (2) Water at 100° C. is changed to steam at 100° C.
> (3) Steam at 100° C. is changed to water at 100° C.
> (4) Frozen alcohol at −130° C. is changed to liquid alcohol at −130° C.
> (Answered correctly by 34 per cent of the pupils.)

It will be noted that these items progressively call for more and more thorough understanding of the heat of fusion of ice. Item 1 requires only a verbal association between "heat of fusion of ice" and "80 calories." This is the sort of association upon which physics pupils are frequently drilled in a more or less mechanical fashion until the association is firmly established. The success with which it has been established in this particular case is evidenced by the fact that 75 per cent of the pupils tested gave the correct response to this item.

Item 2 is of essentially the same type as item 1, but employs a different phrasing from the pat form in which the question is usually stated. Even this slight variation in phrasing resulted in a 5 per cent decrease in the number of correct responses.

The ability to supply the correct answer to either item 1 or 2 clearly can be of no functional value unless the pupil has some notion of the meaning of "heat of fusion." The data from item 3, however, indicate that there were many students who could make the verbal association called for in item 1 or 2 who had no adequate understanding of the meaning of this term. (It may be noted that item 3 was scored in a very liberal fashion, and that many responses were accepted as correct that were technically imperfect.)

Any student who really understood the definition provided in response to item 3 should have no difficulty in responding to items 4 and 5. It will be noted, however, that only 35 and 34 per cent, respectively, of the pupils responded correctly to these latter items. It is clearly apparent from these data that items such as 1, 2, and 3 above can provide only an inadequate basis on which to judge the pupils' understanding of the concept taught. Items of the type of 4 and 5 above are definitely superior. It is significant to observe in this connection that out of the 224 pupils who supplied the correct answer to item 1, only 16 per cent succeeded in all of the remaining items; in other words, only one out of every six students who had acquired the verbal association between "heat of fusion of ice" and "80 calories" had acquired even the low level of understanding of these terms called for in items 2, 3, 4, and 5.

Several factors have probably contributed to the presence in tests of verbal stereotypes that permit successful response by rote. In the first place, test items tend to reflect the character of the instruction that preceded them. Education has been criticized repeatedly, and with considerable justification for its preoccupation with verbal symbols and its neglect of the phenomena to which they refer.

In the second place, it is much easier to build items that hold a pupil responsible for verbal associations than to build items that probe the pupil's understanding. Many item writers have borrowed statements from texts or references, simply rearranging the words, or introducing slight modifications, to produce test items. Almost invariably such items can be answered successfully by anyone who has read the texts or references with some care and who possesses a good memory for verbal associations.

The existence of this common fault is not recognized by many item writers, for it is unlikely that they would set out deliberately to test for rote-learning. But they apparently have failed to appreciate the extreme ease with which a stereotyped phrase, repeated in text or lectures or simply in common speech, is accepted and used by many who have only a vague idea of its original significance.

Items of this type are harmful in two ways. In the first place, they do not provide a valid measure of the desired outcomes of instruction. Such items do not discriminate between those who understand and those who do not. The usefulness of purely verbal associations is strictly limited. Tests emphasizing such associations, if administered at the beginning and end of a course of instruction, may give the impression of striking progress on the part of the student when, in reality, permanent achievements are negligible. In the second place, such items exert an undesirable influence on teaching and learning. Concentrating the attention of teacher and pupils on word memory may crowd out efforts to develop understanding.

The solution to this problem is to be found in better phrasing of items. Practical problem situations and fresh wording of essential ideas should be consciously sought. Efforts may be made to penalize the rote-learner by including attractive verbal stereotypes among the distracters, and by avoiding them as much as possible in the answers.

11. *Avoid irrelevant sources of difficulty.*

Just as it is possible inadvertently to incorporate clues to a correct response, it is also possible to place unintended obstacles in the way of the examinee. Quite frequently reasoning problems in mathematics are answered incorrectly by examinees who have reasoned correctly, but who have slipped in their computations. The following item was designed to measure the principle of price discounts.

1. Mr. Walters was given a 12½% discount when he bought a desk whose list price was $119.75. How much did he have to pay for the desk?

A number of examinees who understood the principle missed the item because of errors in multiplication and in placement of decimal points. If the item is revised as follows:

2. Mr. Walters was given a 10% discount when he bought a desk whose list price was $100. How much did he have to pay for the desk?

the computational difficulty is almost entirely removed so that the principle alone can be tested. Test constructors may differ in their opinions concerning the advisability of eliminating computational difficulty from

certain mathematics problems, but when complex or time-consuming calculations are included in the item, the item writer should recognize that he is chiefly testing computational skill rather than understanding of mathematical principles.

Short-Answer Form

1. *Use the short-answer form only for questions that can be answered by a unique word, phrase, or number.*

The need for this restriction was discussed in the previous section dealing with the characteristics of the short-answer form. The implication of this restriction is that a test composed exclusively of short-answer items is almost certain to overemphasize vocabulary. It is probably safe to observe that written tests in general, both essay and objective, have always placed too great a premium on vocabulary and too little upon other important aspects of achievement.

2. *Do not borrow statements verbatim from context and attempt to use them as short-answer items.*

Ambiguity of the item and perplexing variations in the answers are almost certain to result from this procedure.

3. *Make the question, or the directions, explicit.*

Avoid such indefinite questions as

a. Who was George Washington?
b. Where did Columbus land?

In computational problems, specify the degree of precision expected and indicate whether or not units of measurement must accompany numerical answers.

4. *Allow sufficient space for pupil answers, and arrange the spaces for convenience in scoring.*

It is frequently convenient to have all the blanks in a single column at either the left or the right margin of the examination paper.

5. *In computational problems specify the degree of precision expected, or, better still, arrange the problems to come out even except where ability to handle fractions and decimals is one of the points being tested.*

If the correctness of a numerical response depends upon stating the unit

of measurement, make this fact clear. If not, it is best to include the unit of measurement in the statement of the question as, for example,

c. The volume of a cube nine feet on an edge is ____ cubic feet.

6. *Avoid overmultilation of completion exercises.*

An extreme example of this may be observed in the following sample item.

A hay ______ affords another ______ of the ______ existing between ______ and ______.

A student with a good memory who had encountered this statement before, might be able to puzzle it out and successfully fill the blanks with the words "infusion," "illustration," "relationship," "animals," and "plants." But it is obvious that far too many words have been removed to permit the item to pose a clear-cut problem. Even an expert biologist would find the item troublesome, and he would certainly brand it as trivial (unless he wrote it himself).

The True-False Form

1. *Base true-false items only on statements which are true or false without qualifications.*

Item writers frequently have used broad generalizations and other declarative statements as true-false items. While most true-false items are declarative statements, not all such statements make acceptable true-false items, for many of them involve exceptions and hence should be marked false, if *truth* is interpreted strictly, although they are true in general.

2. *Avoid the use of long and involved statements with many qualifying phrases.*

The difficulty with such statements is that the examinee has trouble identifying the crucial element in the item. If it is necessary to use many words in describing a complex situation for a true-false item, separate sentences should be used. The issue to be judged true or false should be set apart at the end of the item.

3. *Avoid the use of sentences borrowed from texts or other sources as true-false items.*

Very few textbook statements, when isolated from context, are completely and absolutely true. Moreover, many of them are of value chiefly as supporting or clarifying material and are not in themselves highly

significant. Finally, in many cases the meaning of an isolated sentence is not clear.

The difficulties involved in borrowing statements for use as true-false items may be illustrated in the following examples.

1. World War II was fought in Europe and the Far East.
2. A remarkable transaction occurred toward the end of the reign of Constantine the Great.
3. Colloids are near-solutions.

Item 1 is true so far as it goes, but is not completely true, for it fails to mention Africa, the Atlantic, and other battle areas. Item 2 is clearly introductory and has no inherent significance. The meaning of item 3 is not clear enough to permit a decision of "true" or "false." It is unfortunate that many similar statements, lacking absolute truth, basic significance, or clear meaning, have found their way into true-false tests.

Multiple-Choice Form

1. *Use either a direct question or an incomplete statement as the item stem.*

There are some item ideas that can be expressed more simply and clearly in the form of incomplete statements than in the form of direct questions.

1. The present Russian government is a
   a. democracy
   b. constitutional monarchy
   c. Communist dictatorship
   d. Fascist dictatorship

If this item were written with a direct question as the stem it would require more words and read less smoothly.

2. The present Russian government is of which of the following types?
   a. A democracy
   b. A constitutional monarchy
   c. A Communist dictatorship
   d. A Fascist dictatorship

On the other hand, some items require direct question stems for most effective expression.

3. What part are physical scientists playing in public affairs at present?
   a. They concentrate on science and leave public affairs to others.
   b. They are working on a new form of government based on scientific principles.
   c. They are taking greater interest in political and social questions than ever before.

d. They have assumed positions of leadership and control in governments all over the world.

The incomplete statement is less clear and direct.

4. The present part played by scientists in relation to public affairs is one of
   a. withdrawal from active participation in order to concentrate on science.
   b. work on a new form of government based on scientific principles.
   c. greater interest than ever before in political and social questions.
   d. leadership and control in governments all over the world.

At present there is no experimental evidence on the relative efficiency of the two types of stem. Some experienced item writers exhibit a strong preference for the direct question. Others prefer the incomplete statement. Probably the effect of stem type upon the quality of an item is not large. *There are, however, indications that beginners tend to produce fewer technically weak items when they try to use direct questions than when they use the incomplete statement approach.* Several reasons for this tendency may be suggested.

First, because of its specificity the direct question induces the item writer to produce more specific and homogeneous responses. When an incomplete statement is used as the item stem, the writer's point of view may shift as he writes successive responses. This tends to confuse the examinee concerning the real point of the item.

Second, it is usually easier for the item writer to express complex ideas (those requiring qualifying statements) in complete question form. The necessity of having the completion come at the end of an incomplete statement restricts the item writer. He is not free to arrange phrases or words to produce the clearest possible statement.

Third, and most important of all, the writer of a direct question usually states more explicitly the basis on which the correct response is to be chosen. Contrast the two item stems below.

5. In comparing the exports and imports of the United States, we find that:
6. In the United States, how does the value of exports compare with that of imports?

Item 6 obviously sets up a much more definite basis for choosing a correct response than the first. The difference here is not inherent in the form, since item 5 could be improved without changing it to a direct question. However, there is a greater tendency for item writers to be vague when using incomplete statements than when using direct questions. Some incomplete item stems are altogether too incomplete, as in the following example.

7. Merchants and middlemen
   a. make their living off producers and consumers, and are, therefore, nonproducers.
   b. are regulators and determiners of price and, therefore, are producers.
   c. are producers in that they aid in the distribution of goods and bring the producer and the consumer together.
   d. are producers in that they assist in the circulation of money.

Restatement of this item using a direct question increases the number of words, but makes it much easier to understand.

8. Should merchants and middlemen be classified as producers or non-producers? Why?
   a. As nonproducers, because they make their living off producers and consumers.
   b. As producers, because they are regulators and determiners of price.
   c. As producers, because they aid in the distribution of goods and bring producer and consumer together.
   d. As producers, because they assist in the circulation of money.

2. *In general, include in the stem any words that must otherwise be repeated in each response.*

The following item is presented in two forms to illustrate this point.

1. The members of the board of directors of a corporation are usually chosen by which of these?
   a. The bondholders of the corporation.
   b. The stockholders of the corporation.
   c. The president of the corporation.
   d. The employees of the corporation.

2. Which persons associated with a corporation usually choose its directors?
   a. Bondholders
   b. Stockholders
   c. Officials
   d. Employees

It is not always possible, or desirable, however, to eliminate all words common to the responses. In a preceding example, dealing with the activities of merchants and middlemen, it was necessary to introduce each response with the word "as" to make grammatical sense. If the retention of common words in all of the responses makes the item easier to understand, they should be retained. In most cases, however, it will be found that the common words can be transferred to the stem without loss of clarity.

3. *If possible, avoid a negatively stated item stem.*

Experience indicates that this approach is likely to confuse the examinee. He is accustomed to selecting a *correct* response and finds it difficult to remember, in a particular isolated instance, to choose an *incorrect* response. The negative approach and the difficulty it frequently causes may be apparent in the following sample items.

1. Which of these is *not* one of the purposes of Russia in consolidating the Communist party organization throughout Eastern Europe?
   a. To balance the influence of the Western democracies.
   b. To bolster her economic position.
   c. To improve Russian-American relations.
   d. To improve her political bargaining position.

2. Which of these is *not* true of a virus?
   a. It is composed of very large living cells.
   b. It can reproduce itself.
   c. It can live only in plants and animal cells.
   d. It can cause disease.

The use of a *negative* approach can sometimes be avoided by rewording the item, by reducing the number of responses, or both. Where use of negatively stated items appears to constitute the only satisfactory approach, they should be grouped together, under special directions to the examinee. Underlining, italicizing, or otherwise emphasizing the "not" is also essential.

4. *Provide a response that competent critics can agree on as the best.*

The correct response to a multiple-choice item must be determinate. While this requirement is obvious, it is not always easy to fulfill. Sometimes through lack of information but more often through failure to consider all circumstances, writers produce items that confuse and divide even competent authorities. For example, experts disagreed sharply over the best response to each of the following questions.

1. What is the chief difference in research work between colleges and industrial firms?
   a. Colleges do much research, industrial firms little.
   b. Colleges are more concerned with basic research, industrial firms with applications.
   c. Colleges lack the well-equipped laboratories which industrial firms maintain.
   d. Colleges publish results, while industrial firms keep their findings secret.

2. What is the chief obstacle to free exchange of scientific information between scientists in different countries?
   a. The information is printed in different languages.
   b. The scientists wish to keep the information secret for their own use.
   c. Scientists do not wish to use second-hand information from other countries.
   d. Countries wish to keep some of the information secret for use in time of war.

The most obvious remedy for this type of weakness is to have the items carefully reviewed by competent authorities. Items on which the experts cannot agree in selecting a best response should be revised or discarded. Expert reviewers may frequently suggest desirable improvements in the wording of the item, but the item writer should not feel bound to accept these suggestions if they do not affect choice of the answer. Some suggested changes may actually weaken the item. Expert reviewers have a tendency to "split hairs at the Ph.D. level," and to prefer the technical jargon and stereotypes with which they are most familiar. The changes in wording they suggest may sometimes make the item more verbose and confusing to the examinees for whom it is intended, or may destroy its ability to discriminate those who understand from those who simply possess verbal facility.

5. *Make all the responses appropriate to the item stem.*

Writers sometimes produce items in which none of the responses is reasonably correct.

1. Loss due to hail is greatest in which of the following cases?
   a. To livestock
   b. To skylights
   c. To growing crops
2. Why do living organisms need oxygen?
   a. Purification of the blood
   b. Oxidation of wastes
   c. Release of energy
   d. Assimilation of foods
3. What process is exactly the opposite of photosynthesis?
   a. Digestion
   b. Respiration
   c. Assimilation
   d. Catabolism

The responses to the first item are not "cases." The responses to the second item are not stated as reasons, as required by the stem. The third item

illustrates a different type of difficulty. It asks a question which has no possible correct answer, since no process is *exactly* the opposite of photosynthesis. In all three cases the items can be improved by rewording.

4. The greatest loss in hail storms for the country as a whole results from damage to
   a. livestock
   b. skylights
   c. growing crops

5. Why do living organisms need oxygen?
   a. To purify the blood
   b. To oxidize waste
   c. To release energy
   d. To assimilate food

6. What process is most nearly the opposite of photosynthesis chemically?
   a. Digestion
   b. Respiration
   c. Assimilation
   d. Catabolism

One fairly obvious indication of inappropriate or carelessly written responses is lack of parallelism in grammatical structure. This is illustrated in the following item.

7. What would do most to advance the application of atomic discoveries to medicine?
   a. Standardized techniques for treatment of patients.
   b. Train the average doctor to apply radioactive treatments.
   c. Reducing radioactive therapy to a routine procedure.
   d. Establish hospitals staffed by highly trained radioactive therapy specialists.

The responses to a multiple-choice item should always be expressed in parallel form. Sometimes this can be achieved by a simple change in wording. In other cases it requires substitution of a more appropriate response. The revised and improved item is given below:

8. What would do most to advance the application of atomic discoveries to medicine?
   a. Development of standardized techniques for treatment of patients.
   b. Training of the average doctor in application of radioactive treatments.
   c. Removal of restriction on the use of radioactive substances.
   d. Addition of trained radioactive therapy specialists to hospital staffs.

6. *Make all distracters plausible and attractive to examinees who lack the information or ability tested by the item.*

In addition to inappropriate distracters resulting from careless writing, there are others resulting from failure to consider plausibility. Consider the following item.

7. Which element has been most influential in recent textile developments?
   a. Scientific research
   b. Psychological change
   c. Convention
   d. Advertising promotion

Only the first response is plausible as an answer to question 7. Another example is provided by item 8.

8. Why is physical education a vital part of general education?
   a. It guarantees good health.
   b. It provides good disciplinary training.
   c. It balances mental, social, and physical activities.
   d. It provides needed strenuous physical exercise.

The alert examinee would reason that nothing can *guarantee* good health, that *disciplinary training* is now in low repute educationally, and that *strenuous* physical exercise is seldom recommended. Such an item might function well as a test of understanding of verbal meaning, but it would not discriminate between those who do and those who do not understand the place of physical education in general education.

Each distracter should be designed specifically to attract those examinees who have certain common misconceptions or who tend to make certain common errors. The mathematics test item which follows illustrates this point:

9. The ratio of 25 cents to 5 dollars is
   a. 1/20
   b. 1/5
   c. 5/1
   d. 20/1
   e. none of these

The examinee who carelessly overlooks the distinction between cents and dollars, or inverts the ratio, will arrive at one of the distracters rather than at the answer. Some item writers have found it helpful to first present multiple-choice item stems as free-response items, and then to use incorrect responses of some examinees as attractive distracters.

7. *Avoid highly technical distracters.*

Item writers, needing additional distracters, are sometimes tempted to

insert a response the meaning or applicability of which is completely beyond the ability of the examinee to understand.

1. Electric shock is most commonly administered in the treatment of
   a. rheumatism
   b. paralysis
   c. insanity
   d. erythema

The first three suggested responses are fairly common terms. The fourth is almost never encountered. It is definitely a "space filler" in this item, but it presents a frustrating problem to the examinee, since he is forced to choose a best answer without knowing the meaning of one of the answers. The level of information or ability required to reject a wrong response should be no higher than the level of ability required to select a correct response. When this is true, an examinee may sometimes arrive at his choice by successively eliminating incorrect answers. Response by elimination has been criticized, but it has one possible advantage over response by direct selection. The examinee may need more pertinent information to eliminate three plausible distracters than he would need to select one correct verbal stereotype.

8. *Avoid responses that overlap or include each other.*

An example of this defect is provided by the following item.

1. What percent of the total loss due to hail is the loss to growing crops?
   a. Less than 20%
   b. Less than 30%
   c. More than 50%
   d. More than 95%

This item is, in effect, a two-response item. The choice lies between responses *b* and *c*. For if *a* is correct then *b* is also correct, and if *d* is correct then *c* is also correct. More subtle examples of this defect are occasionally encountered in item writing.

9. *Use "none of these" as a response only in items to which an absolutely correct answer can be given; use it as an obvious answer several times early in the test but use it sparingly thereafter; and avoid using it as the answer to items in which it may cover a large number of incorrect responses.*

"None of these" is quite appropriate as a response to the correct answer variety of multiple-choice item. It is inappropriate in best-answer items. An examinee may properly reject all suggested responses if he is working under instructions to choose only completely correct answers. He cannot

reasonably be asked to mark "none of these" when his general instruction is to pick the *best* of several admittedly imperfect responses.

"None of these" is a useful response for items in arithmetic, spelling, punctuation, and similar fields where conventions of correctness can be applied rigorously. It provides an easy-to-write fourth or fifth response when one is needed and may be more plausible than any other that can be found. It sometimes enables the item writer to avoid stating an answer that is too obviously correct.

Two dangers are connected with the use of "none of these." The first is that it may not be seriously considered as a possible answer. The second is that the examinee who choses "none of these" as the correct response may be given credit for a wrong answer. To avoid the first danger, the examinee must be convinced at the beginning of the test that "none of these" is likely to be the answer to some items. This can be achieved by using it as the correct response to several easy items early in the test.

The second danger can be avoided by sparing use of "none of these" *as the correct answer* after the beginning of the test, and by limiting its use as the answer to items in which the possible incorrect responses are relatively few. "None of these" would be an appropriate answer to the following item.

1. What is the area of a right triangle whose sides adjacent to the right angle are 3 inches and 4 inches long respectively?
    a. 7
    b. 12
    c. None of these (answer)

Some examinees may miss this item by simply adding 3 and 4. Others might multiply 3 by 4 and forget division by 2. Still others might add 3, 4, and 5. Since the number of possible incorrect responses to this item is limited, they may all be included as distracters, so that only examinees who solve the problem correctly will be likely to choose "none of these."

This situation does not prevail in the following item.

2. What is the sum of	37,859	a. 176,216
	46,212	b. 186,226 (answer)
	39,843	c. 183,127
	62,312	d. None of these

It is obviously impossible to anticipate all of the possible errors students might make in responding to this item. Hence, it would be undesirable to use "none of these" as the answer with only a few of the possible incorrect responses listed as distracters. It is far more appropriate to use it as a distracter.

10. *Arrange the responses in logical order, if one exists, but avoid consistent preference for any particular response position.*

Where the responses consist of numbers, they should ordinarily be put in ascending or descending order. If the responses are small numbers such as 1, 2, 3, 4, or 5, the 1 should occur in the first position, 2 in the second position, and so on. If this is not done, there will be a strong tendency for the examinees to confuse the absolute value of the answer with the response position used to indicate it. Even when the positions are lettered, the examinee may think of them numerically, and indicate a numerical response of 3 in the *c* position even though some other letter is used to represent 3 in the item.

If an item contains one or more pairs of responses dealing with the same concept, these should usually be placed together. In the following item, it is preferable to arrange the responses as shown, rather than to distribute them at random among the choice positions.

8. Which of these would you expect to be anti-inflationary in the United States?
   a. Increased consumption of goods.
   b. Increased exports to Europe.
   c. Limitation of credit to consumers.
   d. Limitation of the size of savings accounts.

In many items, however, there is no objection to assigning the responses at random to the response positions. This gives the item writer an opportunity to balance roughly the number of answers occurring in each position.

Some writers have advocated that obvious answers should be placed in last so that the examinee will be forced to read and consider the distracters before seeing the correct response. There is no evidence concerning the effectiveness of this procedure, and it appears to be of doubtful value. If the answer is so obvious that the examinee will choose it the moment he sees it, placing it last is not likely to help the item much.

11. *If the item deals with the definition of a term, it is often preferable to include the term in the stem and present alternative definitions in the responses.*

The reason for this suggestion is that it usually provides more opportunities for attractive distracters and tends to reduce the opportunity for correct response by verbal association. Consider these illustrations.

1. What name is given to the group of complex organic compounds that occur in small quantities in natural foods and are essential to normal nutrition?

a. Nutrients
b. Calories
c. Vitamins
d. Minerals

2. What is a vitamin?
   a. A complex substance necessary for normal animal development, which is found in small quantities in certain foods.
   b. A complex substance prepared in biological laboratories to improve the nutrient qualities of ordinary foods.
   c. A substance extracted from ordinary foods, which is useful in destroying disease germs in the body.
   d. A highly concentrated form of food energy, which should be used only on a doctor's prescription.

In the second item it is clear that more of the common misconceptions about the meaning of the term "vitamin" can be suggested and made attractive to the superficial learner.

12. *Do not present a collection of true-false statements as a multiple-choice item.*

Such items usually reveal the item writer's failure to identify or specify a single problem. In some cases, the true-false statements are grouped about a single problem and could be easily reworded to make that problem specific. In other cases, the statements are so loosely related that they hardly constitute a single problem at all. This situation is illustrated by the following item.

1. What does physiology teach?
   a. The development of a vital organ is dependent upon muscular activity.
   b. Strength is independent of muscle size.
   c. The mind and body are not influenced by each other.
   d. Work is not exercise.

Here two of the responses show some similarity. The other two are quite diverse. Grouping all in a single item leads the examinee to look for a common principle. It is difficult for him to arrive at any rational basis for selecting a best response. One beneficial change would be to replace responses *c* and *d* by others dealing with muscles or muscle activity, and to reword the stem to point toward this problem.

Matching Exercises

1. *Group only homogeneous premises and homogeneous responses in a single matching item.*

The premises and responses in the following item are not homogeneous.

1. A drawing tool used primarily as a guide to draw horizontal lines		a. dividers
2. Avoidance of erasures, blots, uneven lines, or poorly shaped letters		b. arc
3. Any part of the circumference of a circle		c. French curve
		d. neatness
		e. T-square

Such an item measures only very superficial verbal associations. It is easily solved by those who have only vague concepts.

2. *Use relatively short lists of responses.*

Seldom should more than five alternative responses be suggested for a given group of premises. Two reasons for this recommendation concern the item writer. It is difficult to maintain homogeneity in a long list of responses. Further, long lists of responses reflect concentration on one aspect of achievement which prevents proper distribution of emphasis on all aspects. The other reason concerns the examinee. With few responses, little time need be wasted in hunting for a proper response. The examinee may even fix the responses in mind so that his only problem is that of reading each premise and deciding which response best applies to it.

The only necessary limitation to the number of premises is imposed by the requirement that they must all belong with the same homogeneous group of responses. It is often impossible to find a large number of premises to which the same group of responses constitute plausible matchings. Even where it is possible, the item writer should probably use short lists of homogeneous premises of the same kind, so that he can sample more different topical areas.

3. *Arrange premises and responses for maximum clarity and convenience to the examinee.*

In general it is desirable to use the longer, more complex statements as premises, to arrange them at the left, and to number them as independent items. The responses should be arranged in order, if any logical basis for order exists, to simplify the task of matching.

4. *The directions should clearly explain the intended basis for matching.*

In simple matching exercises, the basis may be almost self-evident, but it should be made explicit in the directions. For classification-type items, specific instructions are needed. Illustrative items presented in the section on item forms show this detail in directions.[3]

[3] See p. 201.

5. *Do not attempt to provide perfect one-to-one matching between premises and responses.*

The same response may be used for more than one premise. Occasionally responses that fit none of the premises should be included. Nothing is gained by attempting to provide equal numbers of premises and responses and to assure perfect matching. On the contrary, something is lost because the examinee may be given an irrelevant clue to one correct response.

The Interpretive Test Exercise

The interpretive test exercise represents a relatively new and highly promising development. Because it constitutes a larger unit than the typical test item, and because it presents special possibilities and problems, it is discussed here as a separate unit.

The interpretive test exercise consists of an introductory selection of material followed by a series of questions calling for various interpretations. The material to be interpreted may be a selection of almost any type of writing (news, fiction, science, poetry, etc.), a table, map, chart, diagram, or illustration; the description of an experiment or of a legal problem; even a baseball box score or a portion of a musical composition. The questions on this material may be based on explicit statements in the material, on inferences, explanations, generalizations, conclusions, criticisms, and on many other interpretations. Since the interpretive exercise may employ any of several item forms, since it includes introductory materials as well, and since it has special possibilities for measurement, it deserves separate discussion.

The following two illustrations suggest the general form and content of the interpretive test exercise, although they by no means represent all of its possible varieties.

> It has been stated that "like Hellas, the Swiss Land was born divided," and also that "political solidarity had a hard, slow birth in the mountains." Certainly the physical features of the Swiss lands in serving sharply to confine movement and widely to separate settled areas did not facilitate intercourse and thus political cooperation. In mountainous Switzerland, at any rate, village communes tended to occupy the narrow lateral valleys of the Alps, where they engaged in agricultural and pastoral pursuits in a state of almost complete political and economic isolation and self-sufficiency. On the other hand, the geographical position of the Swiss lands was such as to induce a continual current of traffic *en route* for the passes of the Central Alps, whilst the major valleys of the principal rivers formed the main highways of communication. Moreover, the Swiss plateau stretching between Lake Constance and Geneva and cupped between the mountain ranges

formed a broad belt of well-watered and relatively low-lying land which was capable of supporting a population much denser than that of the mountains. Actually, it was not the more-favored plateau lands, but certain cantons of the mountains which provided both the leadership in the wars for independence and the nuclear region around which the state grew. The reason seems to be that in the mountain valleys the peasant and shepherd population tenaciously defended their freedom from the encroachment of the feudal powers and largely escaped being reduced to serfdom, as were the inhabitants of the central plateau.

1. What is meant by "the Swiss land was born divided"?
   1) There were many different religious sects.
   2) Different languages were spoken in different parts of the country.
   3) The mountains isolated the people in different parts of the country.
   4) The people fought among themselves.

2. With which of the following does the writer compare Switzerland?
   1) Ancient Rome
   2) Ancient Greece
   3) Medieval Italy
   4) Medieval France

3. Who took the lead in making Switzerland into a united nation?
   1) The traders
   2) The mountain people
   3) The farmers of the plains
   4) The serfs

4. What does the writer try to do in this paragraph?
   1) To describe the factors in the early commercial development of Switzerland.
   2) To point out why Switzerland can never become a united country.
   3) To explain how trade changed the character of the Swiss nation.
   4) To show how geographical conditions affected the political unification of Switzerland.

5. Which of the following is the most appropriate heading for this paragraph?
   1) Early Swiss Commerce
   2) Trade Routes and Their Effect on Switzerland
   3) Geography and Swiss Freedom
   4) Agriculture on the Swiss Plateau

Presidential Electoral Votes in United States
by Political Parties
1904–1944

Year	Republican	Democratic	Progressive
1904	336	140	
1908	321	162	

1912	8	435	88
1916	254	277	
1920	404	127	
1924	382	136	13
1928	444	87	
1932	59	472	
1936	8	523	
1940	82	449	
1944	99	432	

1. Which party held the presidency during 1926?
   1) Republican
   2) Democratic
   3) Progressive
   4) The table does not tell

2. In what year was the Republican victory the most decisive?
   1) 1904
   2) 1924
   3) 1928
   4) 1936

3. Which of these statements about Democratic party strength is supported by the table?
   1) The Democrats won easy victories in both 1912 and 1916.
   2) The Democrats have been by far the strongest political party since 1904.
   3) Democratic party strength has been slowly increasing since 1932.
   4) Democratic party strength has been slowly decreasing since 1936.

4. Between which two consecutive elections was there the greatest increase in the number of Democratic electoral votes?
   1) 1908 and 1912
   2) 1912 and 1916
   3) 1928 and 1932
   4) 1932 and 1936

5. The percentage of the electoral votes received by the Democrats was the *largest* in what year?
   1) 1944
   2) 1936
   3) 1928
   4) 1912

Characteristics

1. *The interpretive exercise provides an opportunity for measuring directly one of the important outcomes of instruction—the ability to interpret and evaluate printed materials.*

Throughout this chapter and in preceding chapters the desirability, as

well as the difficulty, of measuring the important outcomes of instruction directly has been stressed. Certainly the importance of ability to interpret printed materials is beyond question in modern times. What one needs to know and what one must do are most often presented as printed discussions or directions. These materials must be interpreted and evaluated. Their usefulness depends upon the accuracy and depth of penetration of the interpretations made.

The ability to interpret reading materials assumes even greater importance when general scholastic aptitude is under consideration. This ability is probably more essential than any other to success in most branches of education.

It would be difficult to demonstrate empirically that interpretive exercises do actually measure ability to interpret. But such a demonstration is unnecessary. The tasks set in the questions of an interpretive exercise constitute in themselves an operational definition of interpretation. The usefulness of an interpretive exercise as measure of ability to interpret is limited only by the item writer's competence in selecting material to be interpreted and in formulating problems based on it.

An interesting comparison can be made between essay tests, typical objective tests, and interpretive exercises. The essay test, with which written testing began, asks, in effect, "What can you tell about this subject?" The prevalent types of objective test ask, in effect, "What do you know about this subject?" The interpretive exercise asks, in effect, "What are you able to find out from this material?" or "What are you able to do with it as background?" It is not the purpose of this paragraph to compare the merits of these three test devices. Rather it is to point out that the interpretive exercise occupies a somewhat unique and certainly an important place among the various test instruments.

2. *The interpretive exercise provides an effective setting in which to ask meaningful questions on relatively complex topics.*

In the independent item forms (multiple-choice, true-false, etc.) it is difficult to supply the necessary "raw material" with which an examinee may demonstrate his ability to organize, generalize, or evaluate. Often several paragraphs of material would be required. To include these in the item stem would make it cumbersome and inefficient. Once the material has been prepared and once the examinee has read it, the material provides the basis for not one, but many items. The group of related items which are part of the interpretive exercise make more complete use of these background materials than independent items could.

3. *The interpretive exercise reduces ambiguity by providing a common ground of information for both the item writer and the examinee.*

One of the sources of ambiguity in objective items is the fact that the item writer and the examinee approach the same question from different points of view. In answer to the question, "Where was Paul converted to Christianity?" one examinee wrote, "Acts IV," a perfectly correct answer from his point of view. The introductory material in an interpretive exercise helps to set the stage in more minute detail than is commonly possible in independent items. Where both the item writer and the examinee are working in the same explicit setting, it is much easier for the examinee to comprehend what the item asks, and hence to respond to it to the limit of his information and ability.

4. *The interpretive exercise requires both a general ability to interpret and a specific background of terms, facts, and principles related to the material presented.*

General skill in reading is an important factor in the ability of an examinee to interpret selected materials. Knowledge of the special terms and concepts used and familiarity with the general principles and structure of the specific field with which the material deals are likewise important. Within limits, the relative influence of these two factors is under the control of the item writer. He can include items that call heavily upon general interpretive ability and make almost no demands on specific knowledge. Or he can emphasize specific knowledge and minimize general interpretive ability.

It should be noted, however, that the interpretive exercise cannot, and is not intended to, supply a "pure" measure of informational background. Likewise, when it deals with specialized subject matter, it cannot supply a pure measure of general interpretive ability.

For some purposes, pure measures are very useful. On the other hand, there are many situations in which the combination provided by the interpretive exercise is not only unobjectionable, but even preferable, to independent measures. This is especially true if the item writer recognizes and uses his ability to control the relative influence of the two contributing factors.

5. *The interpretive exercise is well adapted to the evaluation of general levels of educational development for individuals having diverse backgrounds.*

In evaluating the general educational development of an examinee, it is almost useless to ask what he remembers about details of factual con-

tent. What he knows on this level depends not only on how recently he has studied but also on where and by whom he was taught.

It is almost equally useless to ask him to state principles or solve problems as taught in the course, because, while there is more common agreement with respect to principles and methods of problem-solving, these are also acquisitions which grow dim with disuse. Thus, an examinee who is poorly developed educationally but who has just come from a course of instruction may, on a typical subject examination, outscore a more highly developed examinee who has been away from the subject matter for some time.

The interpretive exercise has an important advantage for individuals of the latter type. They are not required to reproduce old learnings. Rather, they are asked to demonstrate that they have acquired a general background of experience, certain attitudes and other values, and familiarity with such useful ways of thinking as analysis, organization, etc. A skillful test constructor using interpretive exercises can measure most of the important outcomes of general education.

6. *The interpretive exercise has wide applications, but its use presents some special problems.*

In the description of the interpretive exercise, mention was made of the wide variety of materials that could be presented for interpretation, and of the correspondingly wide variety of abilities that could be called for by the questions asked. This wide applicability has not been generally recognized, so that the interpretive exercise has not yet assumed the important role to which it is entitled in the field of educational measurement.

On the other hand, some of the difficulties encountered in using interpretive exercises should be clearly recognized. Skilled, experienced item writers find it difficult to construct interpretive exercises of high quality. The selection or construction of suitable materials and the identification of item topics that make maximum use of the inherent possibilities of the material are added problems not encountered in ordinary item writing. The interpretive exercise is relatively time-consuming to administer. An hour of testing with interpretive exercises will produce fewer independent scoring units than the same time devoted to independent test items. Partly because of this time factor and also because of the intrusion of general interpretive ability, the interpretive exercise is not an efficient measure of informational background as such.

Suggestions for Writing

1. *Select the type of material to be interpreted for significance and representativeness.*

Attention should be given to various types of material, and the most suitable should be selected. In many respects the problem of selecting materials to be interpreted is similar to the problem of identifying topics for independent items. The criterions of significance and the distribution of emphasis necessary to produce a good interpretive exercise vary from field to field. They can be determined only by one who is competent in the field.

2. *Write or rewrite the material to be interpreted so as to provide for the desired interpretations and to eliminate nonfunctional portions.*

It is rarely possible to find intact passages of material that are ideally suited for interpretive exercises. Most test passages must either be originally written for the special purpose or developed through thorough revision of existing source materials. It goes without saying that the material should be clearly presented and should conform to the highest standards of form except where deliberate alterations are necessary to provide an opportunity for critical comments or for suggested revisions. Often the addition or elimination of a word or sentence will provide new opportunities for significant questions. The task of producing an interpretive exercise is not simply one of finding suitable material and then writing questions on it. It is rather an integrated task in which preparation of the test exercises and modification of the material go hand in hand.

The first step in constructing a test passage often consists of searching through some materials for a reading selection that seems to contain several promising possibilities for interpretive items. The next step is to construct tentative items that exploit all of the item possibilities which the passage presents. The third step is to *rewrite* the passage so as to eliminate from it anything that does not contribute to the items already built and that is not essential to the continuity of the passage. This condensation may require some modification of the original items, or elimination of certain items entirely. A highly condensed version is thus produced that, with reference to the items already constructed, is far more *efficient* than the original passage, or that yields far more items "per line of passage" or per unit of total testing time.

The next step is to reconsider this condensed version to determine whether, by further rewriting or additions, the basis for additional good

items may be introduced. These changes may again require modification or even deletion of some of the original items. This process of reciprocal revisions and additions to the passages and the set of items continues until the writer feels that he has reached the point beyond which further improvement is not worth the effort it costs. The version finally produced may sometimes have little resemblance to, or contain almost no intact paragraphs or sentences from, the original passage, but in all instances it is a much more efficient test exercise than the original.

The preceding paragraphs describe roughly the manner in which most good interpretive exercises are built. Sometimes, an exceptionally competent test constructor may be able to write from scratch an original selection to satisfy certain predetermined specifications. Even then, the first draft of the passage will ordinarily require many revisions as the items are being constructed. In general, the construction of interpretive test exercises requires a much more *opportunistic* approach than is usually followed in test construction.

3. *Construct the items in either multiple-choice or true-false form, with regard to all suggestions previously made for these forms.*

While the multiple-choice form is generally more adaptable to the questions raised in interpretive exercises, there are situations in which the true-false item is entirely adequate and may even be preferred because of its simplicity.

4. *Decide in advance how much emphasis should be placed upon a student's background information and then construct questions to provide the desired emphasis.*

It has been pointed out previously that the interpretive exercise provides the opportunity for questions that make much or little demand upon the student's background of special information. This possibility should be clearly recognized, and the items written in terms of a definite purpose.

Conclusion

Throughout this chapter attention has been called to the many subtleties involved in item writing, and to the high degree of skill needed by the item writer in dealing with these subtleties. It would be unfortunate if the net effect of these comments were to discourage potential item writers from undertaking the task. Awareness of ideals and high standards need to be balanced by a realistic view of the practical limitations in time and skill which often force the use of substandard tests. Such tests are often far better than no test at all.

In view of the time required to produce good test items, there is a strong incentive for the exchange among writers of good items. A number of suggestions for cooperation along this line have been offered by leaders in the field. It is to be hoped that a cooperative agency for the selection, classification, and distribution of well-written test items may someday be established.

Selected References

1. ADKINS, DOROTHY C. *Construction and Analysis of Achievement Tests.* Washington: Government Printing Office, 1947.
2. ENGLEHART, MAX D. "Unique Types of Achievement Test Exercises," *Psychometrika,* **7**: 103–16, 1942.
3. HAWKES, HERBERT E.; LINDQUIST, E. F.; and MANN, C. R. *The Construction and Use of Achievement Examinations.* Boston: Houghton Mifflin Co., 1936.
4. *How to Make the Picture Test Item.* (DA AGO PRT–873.) Washington: Department of the Army, 1948.
5. MOSIER, CHARLES I.; MYERS, M. C.; and PRICE, HELEN G. "Suggestions for the Construction of Multiple Choice Test Items," *Educational and Psychological Measurement,* **5**: 261–71, 1945.
6. RUCH, G. M. *The Objective or New-Type Examination.* Chicago: Scott, Foresman and Co., 1929.
7. SCATES, DOUGLAS E. "Complexity of Test Items as a Factor in the Validity of Measurement," *Journal of Educational Research,* **30**: 77–92, 1936.
8. UNIVERSITY OF CHICAGO BOARD OF EXAMINATIONS. *Manual of Examination Methods.* 2nd ed. Chicago: University of Chicago Bookstore, 1937.
9. WEITZMAN, E., and MCNAMARA, W. J. "Apt Use of the Inept Choice in Multiple Choice Testings," *Journal of Educational Research,* **39**: 517–22, 1946.

8. The Experimental Tryout of Test Materials[1]

By HERBERT S. CONRAD
U.S. Office of Education

COLLABORATORS: Dorothy C. Adkins, *University of North Carolina:* Frederick B. Davis, *Hunter College;* Edith M. Huddleston, *Educational Testing Service;* William G. Mollenkopf, *Educational Testing Service;* William B. Schrader, *Educational Testing Service*

AFTER A SET OF TEST ITEMS HAS BEEN WRITTEN, CRITICIZED BY SUBJECT-matter experts, and revised on the basis of their criticisms, it must ordinarily be tried out experimentally on a sample of examinees. This sample should be as nearly like the population with whom the final form of the test is to be used as is reasonably possible.

The various purposes which may be served by a tryout are listed below roughly in the order of their practical importance in educational achievement test construction. In practice, not all of these purposes, of course, are served by the tryouts for all tests. Some tryouts are planned with certain of these purposes in mind, others with other purposes.

1. To identify weak or defective items and to reveal needed improvements. More specifically, to identify ambiguous items, indeterminate items, nonfunctioning or implausible distracters, overly difficult or overly easy items, and so forth.
2. To determine the difficulty of each individual item, in order that a selection of items may be made that will show a distribution of item difficulties appropriate to the purpose of the finished test.
3. To determine the discriminating power of each individual item, in order that all items selected may contribute to the central purpose of the finished test and together constitute an efficient measuring instrument.
4. To provide data needed to determine how many items should constitute the finished test.
5. To provide data needed to determine appropriate time limits for the finished test.

[1] This chapter is a condensation and revision of an originally longer presentation. The materials in this chapter do not necessarily reflect the viewpoint or policy of the Office of Education, Federal Security Agency.

6. To discover weaknesses or needed improvements in the mechanics of test taking, in the directions to examiner and examinee, in the provisions for the responses, in the sample or fore-exercises, in the typographical format, and so forth.

7. To determine the intercorrelations among the items, in order to avoid overlap in item selection and to know how best to organize the items into subtests.

The experimental tryout may be simple or it may be extremely elaborate and expensive. The nature of the tryout depends on the purposes for which the test is to be used, the nature of the items, and the amount of time, money, and resources available. In general, the tryout can be relatively simple for sets of items designed to provide parallel forms of tests already in use. In such cases no tryout of general format or of administrative procedures is necessary, and the content and difficulty of the items that have been found satisfactory in previous forms can be duplicated rather readily. If, however, a new kind of test is to be constructed for an important use, an elaborate tryout or series of tryouts must be planned. For example, if novel types of test items were to be utilized for determining eligibility for teaching positions in several large cities, it is obvious that one would want to be very sure of the adequacy of the items before they were administered to the candidates.

Tryout Stages

For purposes of discussion in this chapter, it may be helpful to break the process of experimental tryout into three stages: the pretryout, the tryout proper, and the final trial administration.

Pretryout

Before the items may be tried out, they must, of course, be assembled into tryout units, or tryout forms. By a pretryout is meant the preliminary administration of the tentative tryout units to small samples of examinees for the purpose of discovering gross deficiencies, but with no intention of analyzing pretryout data for individual items. The tentative tryout forms may be "dittoed," or may even be carbon copies of a typewritten draft. The pretryout may be highly informal, and may involve the administration of the tentative tryout units to from half a dozen to a hundred examinees fairly representative of the population to whom the finished test is to be administered. The sample may, in fact, consist merely of a few adults who try to "put themselves in the position" of the students for whom the test is intended. Such a sample is presumably better than none, though it certainly cannot be recommended as the best.

Ordinarily, the test constructor will wish to administer the pretryout himself, since much may be learned by direct observation and by interviews with examinees that would be difficult to transmit in a written report by an assistant. Major omissions, ambiguities, or inadequacies in the directions to the examinee and in the sample items and fore-exercises may be discovered in the pretryout. Serious errors in the assembly of the test booklets or difficulties in the mechanics of test taking may likewise be revealed. The amount of time that should be allowed in the later tryout administrations of the units can be estimated with considerable accuracy, and this is often the major reason for a pretryout. Whether the average level of difficulty of the items is satisfactory can also be quickly determined within rough limits. If the need for a larger number of easy or difficult items is apparent, they can be prepared and incorporated in the revised tryout forms.

Tryout

Once the gross deficiencies in the tryout forms have been eliminated, perhaps on the basis of a pretryout, it becomes necessary to obtain accurate information concerning the performance of each item in a sample of examinees similar to those with whom the final form of the test is to be used. Administration of the tryout forms of the test for this purpose to 400–500 or more examinees is regarded for the purposes of this chapter as constituting the tryout proper. Sometimes, the results of the first tryout reveal the desirability of altering some items and of replacing others. In these circumstances, a second tryout may be needed to ascertain the adequacy of the revised set of items. More than two tryouts may sometimes be deemed advisable, particularly if the tests are to be used to make important decisions about the examinees, and if time and resources are such as to permit getting proof in advance that the final form of the test will be highly efficient as well as valid. However, only very rarely is more than one tryout feasible in practical test construction.

Trial Administration of the Finished Test

On the basis of the data obtained in the tryout, the items are selected and assembled into the finished test. The finished test is then ready for a trial administration.

The trial administration, as it is defined in this chapter, serves to indicate exactly how the test will function in actual use. This means that no material changes can be made after the trial administration, and that the sample employed must be essentially like the group with whom the test is to be

used. The trial administration is thus a "dress rehearsal," and provides a final check on time limits and on the procedure of administration.

This definition means that in many instances the first practical use of the finished test is actually the trial administration.

Planning the Tryout

SAMPLING

Unless tryout data are based on samples of examinees essentially similar to those with whom the test is to be used, the data may be misleading rather than helpful. The indices of difficulty and of discriminating power of the items, the attractiveness of the distracters, and the magnitudes of reliability and validity coefficients for the tryout forms are all dependent on the characteristics of the sample of examinees tested. The test constructor should try not only to obtain representative samples in order to avoid bias in his data, but he should also try to obtain samples that are *efficient* in the sense that they yield maximum information about the population per individual tested.

It is a common mistake to judge the adequacy of a tryout sample solely in terms of the number of *pupils* tested. However, the *school,* as well as the pupil, must be taken into account. Differences in mean test achievement from school to school are usually far larger than would result from simple random sampling, and are often almost as great as differences among individual pupils within a single school. The same item may prove very difficult in one school and very easy in another, or very discriminating in one school and nondiscriminating in another. Even that which is measured by an item may differ from school to school; the same item may prove to be a reasoning item for pupils with one set of learning experiences, and a memory item for those who have had other learning experiences. For these reasons, the number of schools represented in the tryout is important, as well as the number of pupils. A tryout sample of 20,000 pupils all taken from the same school system will generally not serve as well as a sample of 400 pupils taken from many school systems. Theoretically and actually, a sample consisting, for example, of every twentieth child in every sixth grade in a state is vastly better than one of the same size consisting of every sixth-grade child in each one of several schools scattered throughout the state. Articles by Lindquist (*3*) and Marks (*4*) deal with this point.

Practical considerations, of course, play a major part in obtaining samples. Schools and classes within schools constitute natural administrative units, and it is, therefore, ordinarily impossible to test only every tenth pupil

or only every twentieth pupil. This means that if a certain class is to be used at all, every pupil in that class usually has to be tested, even though the resulting sample is much larger than is required. Sometimes, for the purpose of computing tryout data, a random group may be selected from among those tested, thus saving unnecessary clerical labor; but ordinarily a score for each pupil must be obtained for reporting to the cooperating schools.

Administrative Arrangements

In general, the administrative factors to be considered in the tryout of the test are the same as those to be considered in the administration of the finished test in practical use. A thorough treatment of these factors will be presented in chapter 10 "Administering and Scoring the Test," and hence only a brief discussion of administrative arrangements is necessary here. The present discussion is, therefore, to be regarded only as a summary of most important points, and any reader who is seeking help in the planning of an actual tryout should rely primarily on chapter 10 so far as administrative arrangements are concerned.

The most important single principle to follow in planning the tryout is that the tryout materials should be administered under as nearly as possible the same conditions as those under which the finished test will be administered. This principle should be borne in mind in relation to each of the following paragraphs.

Distribution of tryout materials to participating schools.—All materials needed in the schools participating in the tryout should be distributed to them well in advance of the date of the tryout administration, so that adequate opportunity is given for the correction of shipping errors, answering of questions concerning administrative arrangements, and so forth.

Examiners.—Since the care and skill with which tests are administered depend on the test administrators, it is obvious that a good deal of thought should be exercised to secure competent examiners who will conscientiously carry out the directions provided. Ordinarily, the purpose of the testing and the basic rationale of the testing procedures should be explained to each examiner well in advance of the tryout.

Directions to examiner.—The directions supplied to each examiner should be highly detailed and crystal clear. They should define the duties of proctors and examiners and include specific instructions regarding the disposition or return of test materials. If sample exercises of any kind are included in the tryout materials, the examiner should be provided with the answers to such exercises to avoid the embarrassment or the errors that may

occur as a result of haste or unpreparedness when examinees ask for the correct answers. The necessity for adhering rigidly to time limits should naturally be emphasized.

Directions to examinees.—The directions to the examinees in the tryout forms should be as nearly as possible identical with those that are to be used with the finished test. If there is any doubt about the adequacy of the directions, they should be revised on the basis of a pretryout, rather than after the tryout.

There is some disagreement among psychologists and educators, with reference to the regular use of finished tests, as to whether the examinees should be told to guess or not to guess on multiple-choice items for which they are not sure they know the correct answers. To a lesser extent this disagreement also prevails with reference to tryout forms.

Cronbach and others (*1*) have pointed out that the extent to which pupils guess on multiple-choice tests depends not only on the level of their knowledge but also to a considerable extent on personality factors which we do not ordinarily wish to include in our measurement. This is no doubt true, and has been cited as a reason for directing every examinee to mark every item in the tryout form, even though the time allowed is inadequate for careful consideration of each item. It should be pointed out, however, that some pupils react strongly to being forced to mark answers to items to which they have no idea of the correct answer, and some teachers object to having their pupils so directed. Thus, personality factors are apt to influence test scores to some extent whether the directions do or do not require the examinees to answer every item.

Many psychologists recommend directions that ask the pupils to guess only when they have some clues and can make "shrewd guesses"; with such directions, a recommendation is also frequently made in favor of a scoring formula that corrects for chance success. As pointed out in chapter 10, the assumptions underlying the correction for chance success are rarely, if indeed ever, exactly satisfied; hence, this procedure has been criticized, particularly in tryout situations. There is nearly universal agreement, however, that examinees should be told clearly in the directions for a tryout test whether to answer every item or to guess shrewdly and whether the scoring formula includes a penalty for guessing incorrectly.

Sample items.—Consideration should always be given to illustrative or sample items included in the directions. If the items in a test are of a kind familiar to the examinees, sample questions may be superfluous unless a "warm-up" exercise is judged essential. Sample questions or practice exercises may become of special value if it is desirable to control the "set"

of the examinees with regard to speed of response. If an examinee is presented with ten sample items and he finds that he has answered all ten of them in half the time allowed, he learns better than any set of verbal directions could make him learn that he can afford to work more slowly and carefully. Not all test constructors agree that sample items should be used for this purpose. Sometimes, unfortunately, the "set" appropriate for the sample items is entirely inappropriate for the content of the test itself, as examinees have sometimes discovered to their dismay. This situation should be avoided.

Reports on test administration.—Provision should always be made to secure from the tryout test administrators complete reports on the conditions that prevailed during the testing. Full information should be provided regarding any interruptions or irregularities that may have occurred or difficulties that may have been encountered.

Security.—The importance of making certain that every copy of a tryout test has been returned or accounted for varies greatly with the nature and purpose of the test. Sometimes, each copy of a tryout test should be numbered, and test administrators should have to sign receipts for specified numbers of copies. This makes it possible to hold each administrator responsible for returning or accounting for the disposition of every copy in his possession. Sometimes, it is desirable to issue instructions to examiners to inspect each test booklet when it is returned to be sure that all pages remain intact. Other security measures, such as the use of sealed packages, can be employed as needed.

Pupil motivation.—The level of pupil motivation is always an important consideration in the administration of tests, but the problem may be particularly troublesome with tryout tests. If the examinees are told that the results are to be used only for experimental purposes, their motivation may be adversely affected. On the other hand, it may sometimes be unfair to use the results of imperfect tryout forms to judge the pupils' performance. A practical compromise may consist of telling the examinees that the test results will be included in their records for such use as seems justifiable.

Reports to cooperating institutions.—Almost invariably the scores of all examinees should be reported to them or to their teachers within a reasonable period of time. Knowledge that this will be done is an important factor in securing good motivation among the group tested and a high degree of cooperation among the teachers and administrators who have arranged for the testing. Effective means for providing prompt, convenient reports of the results of testing should always be worked out in advance

and adhered to with scrupulous care. Too often, such reports are neglected or are provided only after interest in and uses for the scores have long since disappeared.

Securing reactions of examinees.—Provision should be made for noting down comments by examinees regarding the directions and content of the pretryout and the tryout forms, and such comment should be carefully considered in revising the test. Inadequate directions and faulty items are often identified in this way.

Undesirable attitudes toward schools and especially toward examinations have undoubtedly been engendered in students by correctible faults in examinations or in the administrative procedures connected with them. Unsophisticated test constructors sometimes tend to discount criticisms of examinees (and even of subject-matter authorities) when statistical data indicate no apparent basis for their criticisms. This is an unfortunate tendency because the criteria for item analysis typically leave much to be desired. Although item analysis data are often exceedingly helpful in test construction, they cannot and must not be relied upon to detect all the ambiguities and other faults of items. For various reasons, then, test constructors should accept with proper humility the comments of the examinees.

Scoring procedure.—The scoring procedures for use in each tryout stage and with the final form of a test should be worked out carefully before the tryout forms are reproduced because the requirements of quick and easy scoring sometimes influence the arrangement of items. For example, forethought in planning tryout tests may make possible obtaining separate scores for as many as six groups of items from only one insertion of an answer sheet in the IBM test-scoring machine.

Scoring formulas should be selected to yield maximum validity with minimum labor. They must ordinarily be coordinated with the directions and timing for the test; thus, if the examinees are told to guess, but are given insufficient time to answer all items, a correction for chance may be called for. Various scoring formulas are presented and discussed in chapter 10. Sometimes data from the trial administration should be used to calculate the relative weights for "right" and "wrong" answers that will cause scores to yield maximum correlations with designated criterion variables (*2*, pp. 458–61). Convenient approximations to the optimal weights can then be incorporated in the scoring formulas for the final form of the test.

As stated in chapters 9 and 10, many test technicians recommend correction for chance in scoring. Some test technicians recommend the use of

correction for chance in obtaining the scores to be used as the criterion for internal-consistency item analyses even if the item analysis data for each item are not so corrected and if the final form of the test is not to be scored with correction for chance. This is particularly important if not every examinee in the tryout sample is able to reach nearly every item within the time limit. Correction for chance in obtaining criterion scores may sometimes remove contaminating influences (such as rate of work and some personality differences) from the criterion variable.

Test Materials

Layout.—Careful planning of the arrangement of items within a tryout test must precede reproduction of the copy. Many of the mechanical details pertaining to this problem are discussed in chapter 11. Here, we should consider some more fundamental decisions.

Insuring adequate tryout of all items.—Since one of the purposes of a tryout is to gather accurate data about each individual item, it is extremely important that the time limits be so generous as to permit all, or nearly all, of the examinees to attempt to answer every item on which tryout data are desired. This poses a real problem when the items are to be used in the final form of a speeded test, that is, one so timed that a considerable proportion of examinees do not have time to finish. If the experimental tryout were conducted under speeded conditions, the items toward the end could not be attempted by a large proportion of the examinees; but if the tryout were conducted with generous time limits, the mental-set and rate of work of the examinees would be so different from those likely to obtain during administration of the final form of the test that the tryout data might not be useful.

Four ways of avoiding this dilemma have been widely employed. Under the first method, the items to be tried out under speeded conditions are followed in the test booklet by a set of "cushion" items of the same general type as the others but that are neither to be analyzed nor scored. These items should be relatively difficult and time-consuming, since their only purpose is to keep the faster or abler examinees occupied during the latter part of the test period and thus prevent them from distracting other examinees or creating a disciplinary problem. Under the second method, the items to be tried out are placed in different orders in two or more tryout booklets. That is, if 30 items are to be tried out, items 11–20 in booklet A appear as items 1–10 in booklet B, and items 21–30 in booklet A appear as items 1–10 in booklet C. Thus, every item appears among the first ten in at least one of the tryout booklets, and item analysis data

may be computed only for the first 20 items in each booklet. The second method is essentially the same as the first—the cushion items in each form consisting of some of the items to be tried out. The second method is, of course, the more expensive, since it requires the printing of several tryout forms for the same items, and utilizes only a part of the total tryout sample on each item.

Under the third method, only the items to be tried out constitute the tryout form(s), the items are administered with directions to guess shrewdly but to avoid blind guessing, and the time limit is set so that only a small proportion of the examinees will not have time to finish. An effort is then made to utilize the tryout data for the end items—those not attempted by all examinees—by making various corrections for guessing and for "non-attempts" in the difficulty and discriminating indices for these items. Corrections of this character are presented in chapter 9 following.

The fourth method is especially applicable when items designed for parallel forms of tests administered as part of an organized program are to be tried out. It consists of interspersing new items in current final forms or of preparing special tryout booklets for administration along with current final forms. In the National Teacher Examinations, for example, the vocabulary section originally consisted of 90 items of which 10 (comprising one column on one page) were tryout items that were not scored with the remaining 80. Instead, scores on the 80-item current form served as a criterion variable for item analysis of each of the 10 tryout items. This procedure has the merit of removing the spurious element ordinarily present in internal-consistency item analysis data. By making up 10 different sets of tryout items and running each of them in for one-tenth of the total printing, 100 items were tried out each year to provide an 80-item final form for the following year. This is obviously an extremely expensive procedure, and for that reason is rarely practicable.

Most tests of educational achievement are not highly speeded, but are administered with liberal time limits so as to place the major emphasis on "level" or "power" rather than on rate of work. In general, therefore, the best procedure is to provide *very liberal* time limits for the tryout units, so that nearly all pupils have time to *consider* (not necessarily to mark) *all* items that are being tried out, and then to use one of the first two methods just described to provide for the small proportion of pupils who do not finish the test. In most cases, the first and simpler of these methods should prove adequate.

Provision for criterion measures.—A basic distinction in types of try-

outs is that concerned with the provisions for a criterion in terms of which the individual items are to be judged. In general, the best educational achievement tests themselves constitute the best available definition of the type of behavior being measured, and no external criterion currently exists that measures this behavior better than the score on the test itself (see pages 146–48). The items must then be evaluated against an internal criterion, consisting either of the score on the tryout form itself, or the score on an independently administered test like that in which the items being tried out are to be employed. Suppose, for example, that new items are being tried out for the construction of a new form C of a test for which forms A and B are already available. One possibility is to administer the new items in one or more tryout forms, each form being administered to a different sample, and to use the total score on each tryout form as the criterion in terms of which to compute the discriminating power of each item in that form. The other possibility is to administer the tryout items in one or more tryout forms, each to a different sample, to administer either form A or B to all samples, and then to use the score on form A (or B) as the criterion in terms of which to judge all tryout items.

If the first alternative is employed, it is essential that each tryout form be long enough to yield a *reliable* internal criterion. Sometimes it may be possible to administer *all* items to be tried out in a single tryout form, so that all items are taken by the same examinees. This, of course, insures the maximum reliability in the internal criterion, and also makes it possible, if desired, to compute intercorrelations among all of the items. However, it is frequently difficult to induce cooperating schools to administer large tryout units, or it may be otherwise undesirable to do so, for example, because of fatigue or motivation factors.

If the second alternative is followed, the tryout forms may be of any length, and administrative convenience is usually the major factor in determining their length. Consideration should be given, however, to the possibility that the tryout forms may be so short in comparison with the finished test that the attitude of the examinees toward the tryout test (motivation and fatigue factors) may differ too markedly from that which will characterize the finished test. Furthermore, the use of many short tryout units independently administered to different samples precludes the possibility of securing item intercorrelation data.

A good example of the use of the second of these alternative procedures is provided by the Iowa Basic Skills testing program for grades 3-8. In this program, an entirely new edition of tests has been provided each year, and the tryout of new test materials for the succeeding program has been

made an integral feature of each annual testing program. This has been done by assembling the tryout items each year into self-administering experimental units, each of which constitutes a miniature power test with a time limit of twenty minutes, and each of which corresponds to one of the tests in the regular battery being concurrently administered. These units are distributed in rotation so that every pupil in a class of 50 may be taking a different experimental unit. In a testing program involving 70,000 pupils, it is possible to secure 140 samples of 500 pupils each and, correspondingly, to try out 140 twenty-minute units in these samples. To try out all of these items with a single group of pupils would require over 40 hours of their time. Furthermore, the 500 pupils in each sample are distributed proportionately among all schools participating in the program, so each sample is apt to be highly representative of the entire group and thus similar to all other samples.

The criterion measure for each experimental unit, of course, is the score in the corresponding test in the finished battery being used in the regular program. The criterion, then, is undiluted by weak or defective tryout items, and there is no issue of self-correlation so far as the tryout items are concerned. This method minimizes the difficulty of getting examinees for experimental purposes, eliminates the need for special reports to co-operating institutions, and maximizes the representativeness of the tryout samples. Every effort should be made to employ such methods whenever they are feasible.

When the items to be tried out are organized into a number of tryout forms, each of which is to be administered to a separate sample, it is extremely important that highly comparable samples be employed. This will almost certainly not be true if the different samples are made up of different classes and schools. The only satisfactory procedure, generally, is to administer the various forms simultaneously to different randomly selected pupils (or, better, matched pupils) in the same class or group, so that all classes and schools are proportionally represented in each sample.

Order of difficulty of items.—To avoid any possibility of confusing or discouraging the examinees early in the tryout period, easy items or familiar types should be placed near the beginning of a test, and unfamiliar, unusual, difficult, or particularly complex types of items should be placed at the end of a tryout test.

Surplus items for discard.—It is always necessary to try out more items than will be needed for the finished tests. There is no universally applicable rule regarding the margin for discard that must be allowed when a set of items is tried out. It is safe to say that the margin can be minimized

when the items are constructed to provide another form of an existing test, when the items are factual in content or are concerned with narrow and well-defined skills, when the standard of excellence to be met is not specially high, and when the tryout sample is fairly large and highly representative of the group to which the final form of the test will be administered. For vocabulary items tried out in the National Teacher Examinations, a margin for discard of 20 percent was more than adequate. In fact, it was unusual for any vocabulary item in the tryout test to prove unsatisfactory for use in the final form of the test.

In general, the more complex the items or the mental functions tested by them, the larger the margin for discard must be. The margin must also be increased as the number of separate categories of items is increased. This is a consequence of the fact that the percentage of acceptable items within a given category is subject to sampling fluctuations. Suppose that the outline for a history test calls for four items in the category of "names in the news." Even if in the long run about two-thirds of items of this kind that are tried out prove acceptable, it would not be prudent to try out only six items in this category. By mere chance it might be that only two or three of them would prove acceptable. A smaller margin of discard is adequate in the case of categories in the outline that call for a considerably larger number of items because the effect of chance will be correspondingly reduced.

The principle stated above also applies to item categories in terms of difficulty. Ordinarily, only a small proportion of the items in the final form of a test are intended to be exceptionally easy or difficult. For this reason, we are likely to include only a small proportion of exceptionally easy or difficult items in our tryout test. Thus, we are dealing with categories that include only a few items, and we should allow for a larger margin of discard. One reason for the noticeably high mortality of very difficult items is the greater role of chance in determining responses to these items. There is naturally a greater amount of guessing on such items than on easy items.

The excess of tryout items over items needed for the finished test also depends on the competence of the item writer, or upon the quality of the tryout items. Of two item writers constructing items for the same test, one may produce a much larger proportion of defective items than the other. The individual planning the tryout must, therefore, depend to a considerable extent on his own evaluation of the items in deciding how many to try out.

Organization into subtests.—Because the criterion variable for an in-

ternal-consistency item analysis should ordinarily be as homogeneous as possible, items constructed for an achievement test are sometimes tried out in separate groups, the total score on each group being used as the criterion for obtaining item analysis data for each item within the group. In constructing a test of college physics, for example, one might try out separate groups of items in mechanics, heat, electricity, and so forth. This is usually desirable even though all types of items are later to be intermingled in the finished test and all contribute to a single score. It is obviously important that each group be large enough to yield a reliable criterion score, and that all groups be tried out on highly comparable samples.

Data Obtained through the Tryout

Concerning the mechanics of test taking.—Information regarding the adequacy of *directions* to the examiner or examinee is obtained mainly from the reports of the test administrators in the pretryout and tryout, preferably the former. Sometimes the directions for the pretryout form are so incomplete or misleading as seriously to affect the examinees' scores. Blank answer sheets, indicating failure to mark answers to any items, point either to inadequate directions or to lack of moderately easy items. Failure of examinees to go beyond a certain item may be found to result from lack of directions to turn the page or to go on to the next part, and the like.

Reports from test administrators of the pretryout (and tryout) likewise provide most of the data regarding the suitability of *practice exercises* or sample items. Revision of the latter may be required to improve their length or level of difficulty. Adjustment of the time limits for practice exercises is commonly made on the basis of data obtained in the pretryout.

Examiners and proctors at tryout administrations should observe carefully the use made of *scratch paper,* when provided. Scratch paper should always be collected with the answer sheets if test security is to be assured and should ordinarily be inspected later in order to find what use was made of it. The inspection of test booklets, which is essential if they are to be reused, often reveals a need for scratch paper when it has not been provided (the margins and other spaces being found to have been filled with jottings).

If the advice concerning test format that is provided in chapter 11 is followed with the tryout forms, there will be little need to modify the *format* of the finished test on the basis of tryout data. (When any significant change in format is made, a subsequent tryout in the new format is usually necessary.) The examiners and proctors should always be alert, however, to detect instances where the test exercises have been inconveniently arranged or illegibly reproduced. Sometimes, maps or passages necessary

to answering certain items may be placed in such a way as to require constant turning back and forth of pages. Inconveniences of this sort should be eliminated in final printings.

Concerning time requirements.—In general, a test constructor either (1) builds his test to satisfy certain standards of reliability or validity, or both, or (2) he constructs a test to be administered in a predetermined time limit, such as a forty-minute class period. In the first case, he must first determine from the tryout what is the reliability or validity of the tryout form. Given this information, by means of the formulas for estimating the reliability and validity of lengthened tests, he can estimate roughly how many items like those in the tryout form will be needed to yield the desired reliability or validity. Then, having also determined from the tryout how much time per item on the average is required by the examinees, he can compute what time limit is appropriate for the required number of items (subject to later check in the trial administration). In the second case, in which a predetermined time interval is to be employed, he must again determine from the tryout what is the average time per item that is required, and can then determine how many items similar to those in the tryout can be administered in the allotted time.

In either case the test constructor must determine from the tryout what is the average time per item required by the examinees. In general, the best way to do this is to interrupt the examinees at certain times during the tryout period, and request them to mark in a characteristic way the item on which each is working at that instant, having first directed them to answer all questions in the order in which they are printed. For example, in a thirty-minute tryout period, the examiner may interrupt at the end of fifteen minutes to say, "On your answer sheet, please draw a *circle* around the number of the item on which you are now working." Five minutes later they may be asked to draw a *triangle* around the number of the item on which they are then working, and five minutes later to draw a *square* around it. It will then be easy, later, to compute the average number of items considered in fifteen minutes, or in twenty minutes, and so forth, or to compute any percentile in the distribution of numbers of items considered in any given interval. From such data, the test constructor may readily determine how many items similar to those in the tryout will be completed by any desired proportion of the examinees in any given time limit.

Inasmuch as the finished form consists of selected items from the tryout form, and since the selected items may differ systematically in their time requirements from those discarded, it is often desirable to determine

the final time limits in a trial administration of the finished test, using the procedure just described. Where an external criterion is available, there are still better methods for determining an optimum time limit for the finished test (see pages 336–38). The factors to be considered in determining time limits are more thoroughly discussed in chapter 10.

Concerning reliability and validity.—It is often desirable to compute the coefficients of reliability (and of validity, if possible) for the tryout forms, particularly if the finished test is to satisfy certain standards of reliability (or validity). Appropriate methods of computing these coefficients are presented in chapters 15 and 16. Given the reliability (or validity) coefficient of the tryout forms, appropriate formulas (pages 581 and 608) may be employed to estimate the reliability (or validity) of the finished test, if the number of items is predetermined, or to estimate how many items will be required in the finished test to produce the desired reliability (or validity). Since the finished test will be made up of selected items, which presumably are more reliable and valid than the typical item in the tryout form, these estimates will usually be biased, but in the direction of conservativeness.

Concerning difficulty and discriminating power of individual items.—The major reason for the experimental tryout is to secure quantitative measures or indices of the difficulty and discriminating power of individual items. Appropriate indices and ways of computing them, as well as suggestions for their use in test revision, are presented in the following chapter.

Selected References

1. Cronbach, L. J. "Response Sets and Test Validity," *Educational and Psychological Measurement,* **6**: 475–94, 1946.
2. Guilford, J. P., and Lacey, J. I. (eds.). *Printed Classification Tests.* Washington: Government Printing Office, 1947.
3. Lindquist, E. F. "Sampling in Educational Research," *Journal of Educational Psychology,* **31**: 561–74, 1940.
4. Marks, Elias. "Sampling in the Revision of the Stanford-Binet Scale," *Journal of Educational Psychology,* **44**: 413–34, 1947.

9. Item Selection Techniques

By FREDERICK B. DAVIS
Hunter College

COLLABORATORS: Dorothy C. Adkins, *University of North Carolina;* Herbert S. Conrad, *U.S. Office of Education;* Edward E. Cureton, *University of Tennessee;* William G. Mollenkopf, *Educational Testing Service;* Julian C. Stanley, *George Peabody College for Teachers*

AS INDICATED IN PREVIOUS CHAPTERS, THE USUAL PROCEDURE IN building an objective test is to try out more items than will be needed in the final form. Ingenious statistical techniques to aid in selecting the items for the final form have been developed, but the precise circumstances in which these techniques are useful are not widely understood and they have sometimes been misused. A major emphasis in this chapter will, therefore, be placed on when and how to use them.

By the time the tryout forms for a test have been constructed, much of the opportunity for selecting items has already passed. The first and fundamentally most important steps in selecting the items for almost any test are taken when an outline is constructed to determine its content and the items are prepared to measure the skills and abilities included. This outline should ordinarily be constructed on the basis of the judgment of a group of curriculum authorities and subject-matter experts and should indicate the weight to be given to each topic. The writing and editing of test items call for a rare combination of psychological insight, facility in written expression, and knowledge of the subject. Because item writing and editing are difficult and the mechanical use of item analysis data is relatively easy, there has been an unfortunate tendency to minimize the former and rely heavily on the latter. It cannot be emphasized too strongly that statistical data are at best merely a valuable guide in putting a test together and cannot take the place of scholarship, ingenuity, and painstaking effort on the part of the item writer.

Selecting Items after the Experimental Tryout

The difficulty of individual test items is not an important consideration when items are selected for mastery tests or, under some circumstances, for predictor tests. Mastery tests are not intended to provide scores that

will rank students in terms of their knowledge or ability; rather they are designed to separate students into two groups: those who know certain basic facts, principles, or operations, and those who do not know them. For this reason, the items in a mastery test are so chosen that nearly every pupil who has reached a predetermined level of achievement can answer them all correctly.

Predictor tests are constructed in such a way as to maximize their correlations with specified criterion variables. Ideally, if samples large enough and representative enough could be used, product-moment correlation coefficients among all the items and between each item and the criterion would be computed and items would be selected for use in the final form of the test if they had multiple regression weights significantly different from zero at a specified level of confidence. Under these circumstances, the difficulty of each individual item would not have to be considered separately since it would automatically play its part in determining the item correlations. In actual practice, however, samples large enough to establish the existence of non-chance unique variance in every one or in nearly every one of a set of test items (assuming that it actually exists) cannot be used. Hence, this procedure is rarely defensible, and methods of item selection that involve the separate use of item difficulty indices should be employed in the construction of predictor tests.

However, most educational tests are neither mastery tests nor predictor tests, and the use of item difficulty indices is a matter of considerable importance in their construction. For this reason, a discussion of the various methods used to determine item difficulty is appropriate at this point.

Indices of Item Difficulty

Many ways of expressing the difficulty level of an item have been proposed. The most obvious of these is the percent of the tryout group that marks it correctly. A serious disadvantage of this procedure is that it makes no allowance for the fact that, when the items are of the multiple-choice type, some examinees may guess blindly among the choices presented and mark the correct answer by chance alone. When the number of choices in the item is reduced to the minimum of two, this disadvantage becomes more serious. On a two-choice item, an examinee has a fifty-fifty chance of answering correctly even if he guesses blindly. To cope with this troublesome problem, a correction for chance has been proposed.[1] There is, however, rather sharp disagreement among test technicians re-

[1] For a derivation of the formula commonly used to correct for chance success see "Selected References," p. 325, item 35.

garding its appropriateness. For this reason, both points of view will be presented and the reader may form his own judgment.

Correction for Chance Success in Computing Item Analysis Data

The principal assumption involved in making use of the conventional correction for chance success is that examinees who do not possess enough knowledge to permit them to select the correct answer will guess blindly among all the choices (correct and incorrect) in the item. The extent to which this assumption is satisfied in actual practice varies with the nature of the items and of the groups tested. To the extent that examinees choose incorrect answers on the basis of misinformation rather than on the basis of blind guessing, the procedure overcorrects for chance success; to the extent that they can eliminate from consideration incorrect choices on the basis of partial information that is correct but not adequate to permit identification of the correct choice, the procedure undercorrects for chance success.

Most test constructors agree that few examinees actually guess blindly among all of the choices in any item they have time to consider. Some of the examinees will ordinarily select an incorrect choice in the belief that it is actually correct. They are acting on the basis of misinformation. Others, probably often constituting a majority of those who mark the item incorrectly, have a certain amount of valid information about the point being tested and can eliminate one or more of the incorrect choices from consideration. The more perceptive of them apply every kind of criterion to the remaining choices. "What does the examiner want me to answer? What clues to the right answer are there?" are typical of the questions examinees try to answer in order to mark the correct choice. Finally, a preference among the choices that cannot be eliminated from consideration is often made on the basis of some factor such as the relative lengths of the choices, or the precision of the language in which they are expressed, or the distribution of responses among the choices in the items already answered. In a well-constructed test the relationship between these irrelevant considerations and the correct answers to the items is zero, or even slightly negative. Hence, the examinees will do no better than chance would permit when they make use of these non-chance but irrelevant considerations.

This analysis of the reactions of examinees who cannot identify the answers to test items is based on observation, on introspections reported by examinees, and on certain observable test results. It is, of course, oversimplified; in practice, the factors that underlie the behavior of examinees

are woven into complex patterns. Among the observable test results that offer clues to the reasons underlying the responses to items are the negative percentages sometimes found after correction for chance success and the great differences in the attractiveness of the distracters in most items. Almost any extensive file of item analysis data contains items that are answered correctly by less than the most probable proportion of the examinees that would answer correctly by chance alone. If the number of items answered correctly by less than the most probable proportion of the examinees is significantly larger (at, say, the 5-percent level of confidence) than would appear by chance, this constitutes preliminary evidence that the examinees marked some items largely on the basis of misinformation.

If examinees who are not able to select the correct answer to an item were to guess blindly among all the choices, the incorrect choices would be chosen by insignificantly different proportions of examinees. In this case, the proportion of examinees who marked the correct answer by chance alone would presumably be closely approximated by the average of the proportions marking the incorrect choices. Anyone who has examined item analysis data, however, knows that only rarely are the incorrect choices in an item found to be equally attractive to the examinees. Presumably, this indicates that examinees who do not know the correct answer are responding not merely on the basis of blind guessing but, to some extent, on the basis of partial information and misinformation. Unfortunately, there is no straightforward way of estimating the proportion of the examinees who marked the correct answer by chance alone when (as is usually the case) misinformation as well as partial information and blind guessing plays a part in determining the examinees' responses. If we ignore the effect of misinformation, however, such an estimate can be made. Say, for example, that for a five-choice item 20 percent of the examinees are able to select the correct answer (choice B) on the basis of the information they possess and that 10 percent are so nearly uninformed that they can respond only on the basis of a blind guess (or resort to irrelevant non-chance factors uncorrelated with the function measured by the item). In this situation they will distribute their marks equally among the five choices in accordance with chance expectancy. Another 20 percent of the examinees are sufficiently informed to be able to eliminate choice E but cannot discriminate among the other choices; so they distribute their responses equally among choices A, B, C, and D. Another 18 percent are able to eliminate choices A and E, and they distribute their responses evenly over choices B, C, and D. The remaining 32 percent are able to eliminate choices A, E, and C but do not possess enough information to discrimi-

nate between choices B and D. Hence, they split their responses between these two choices. The item analysis will then show the following distribution of responses:

	Percent
Choice A	7
Choice B	49
Choice C	13
Choice D	29
Choice E	2
Total	100

Now what part of the 49 percent who marked the item correctly did so by chance? We postulated that only 20 percent knew the answer; therefore, 29 percent must have marked choice B by chance alone. Note that this is exactly the percentage that marked the most popular incorrect choice. That these two percentages are identical is not a coincidence. Horst has shown mathematically that in the case of a multiple-choice item where the correct answer is the most popular choice and the choices constitute a graded series of steps of *knowledge* about the point being tested, the percent of the examinees selecting the most popular incorrect choice is equal to the percent marking the correct answer by pure chance. By this restriction Horst rules out misinformation (*78*).

The point of any correction for chance success is to subtract from the percent marking an item correctly the percent attributable merely to chance or to non-chance factors having zero relationship to the function measured.[2] Yet we have shown that the percent to be subtracted cannot be estimated precisely under the conditions that most often obtain.

The conventional correction for chance success leads to the following formula for computing the percent of successes on a given test item that is read by every examinee:

$$R_p - \frac{W_p}{k-1},$$

where R_p = the percent who answer correctly,
W_p = the percent who answer incorrectly,
k = the number of choices in the item.

When this formula is applied to the data for the item given above, we obtain for the corrected percentage:

$$49 - \frac{51}{4} = 49 - 12\tfrac{3}{4} = 36\tfrac{1}{4}.$$

[2] Our discussion of correction for chance success does not consider the problem of assigning optimal weight to error scores when both "rights" and "wrongs" are entered into a multiple regression equation yielding scores maximally correlated with a set of criterion scores.

Note that we subtract only $12\frac{3}{4}$ percent though 29 percent did select choice B by chance alone. We have greatly undercorrected for chance success when we used the conventional correction procedure. This will always be the case when misinformation has played no part in determining the examinees' responses to the item and when the incorrect choices are unequally attractive because of the operation of partial information.

In actual practice, misinformation also operates to determine in some degree the examinees' responses. As will be apparent on a moment's reflection, the presence of misinformation causes the conventional procedure for correcting for chance success to overcorrect. Hence, the two factors other than chance that operate to determine the responses of examinees who cannot identify the correct answer tend to counterbalance each other in their effects on the conventional correction for chance success. Whether partial information or misinformation is the more important in determining the examinees' responses varies from item to item and from group to group of examinees.

In view of the lack of analytic precision in the conventional correction for chance success, it is not surprising that there is some disagreement among test constructors about whether to make use of it in computing item analysis data. Test technicians who oppose its use for this purpose contend that blind guessing on the part of the examinees may be reduced to *negligible* proportions by:

1. warning pupils against it,
2. constructing items in such a way as to include in their distracters bits of misinformation or of partial information about the subject matter that are known to be in common circulation among the examinees,
3. allowing sufficient time for every pupil to consider all the items being tried out.

Whether this contention is correct cannot be completely proved or disproved, but it seems reasonable that these steps would tend to reduce the amount of blind guessing. Let us consider the practical implications of these three suggestions.

Warning pupils against blind guessing can be accomplished in several ways. For example, the directions to examinees might state, "You may answer questions even when you are not sure that your answers are correct, but you should avoid *wild* guessing." This statement affects examinees in different ways. Those who try to follow the admonition must decide how little confidence they can have in an answer before it becomes a wild guess. Naturally, the more conscientious and timid examinees will omit

items more often than will others. Some examinees will deliberately answer all items if they think that the scoring system provides no larger penalty for guessing wrong than for omitting an item. Thus, the bolder, more sophisticated, and less conscientious examinees may enjoy an advantage deriving from their personality characteristics and not at all from their understanding of the subject matter being tested.

Many test constructors and teachers object to this possible outcome and urge that a penalty for errors be included in the scoring system so that examinees who guess freely or who callously ignore the directions will not enjoy this advantage over those who omit items that they feel they cannot answer with at least a shrewd guess. To provide this penalty, the conventional correction for chance success may be employed, and the directions may be modified to read, "You may answer questions even when you are not sure that your answers are correct, but you should avoid *wild* guessing. Your score will be the number of items you mark correctly minus a fraction of the number you mark incorrectly."

Occasionally it has been suggested that these directions be used even when no penalty for errors is employed; more often, it has been suggested that the directions imply a penalty for errors when none is to be used. Experienced test constructors reject these suggestions because the effect on pupils and on subsequent test administration is serious if the ruse is discovered, as it almost invariably will be.

Constructing items in such a way that the distracters will be plausible is universally regarded as desirable, because it tends to prevent examinees with only partial information from eliminating certain choices and thus improving their chances of guessing correctly and because conscientious examinees who lack sufficient knowledge to select the correct answer may find distracters that suggest bits of misinformation in their possession and mark those distracters. In addition, the test constructor should make the correct answers unlike stereotyped responses that some examinees may have memorized without genuine understanding. How successful item writers are in attaining these goals depends on a number of factors, such as their skill and the nature of the subject matter. To the extent that plausibility in distracters is obtained by incorporating popular misinformation into them, blind guessing tends to be reduced; to the extent that plausibility in distracters is obtained by making them only slightly incorrect, the *effectiveness* but not the *amount* of blind guessing tends to be reduced. That is, if all the distracters in an item are so plausible that an examinee's partial information is not adequate to rule out any of them, he will have to guess among *all* of the choices or omit the item. Thus, as

the plausibility of the distracters is increased by making them only slightly incorrect, the conventional correction for chance success becomes more nearly applicable, whereas if the plausibility of distracters is increased by incorporating misinformation into them, the conventional correction for chance success becomes less applicable.

To allow sufficient time for every examinee to consider every item in a tryout test is a worthy objective but one that is harder to attain, in actual practice, than one might at first suppose. Let us consider why it is ordinarily so desirable for every pupil to have time to consider every item. First, the reliability of the item analysis data pertaining to any given item is a function of the number of examinees who mark it; second, the functions measured by items for which it is appropriate to obtain individual item analysis data are not ordinarily speeded functions but are power functions; third, the behavior of examinees who know they are faced with more items than they can seriously consider within the time limit varies considerably. The bolder, less conscientious, and more sophisticated of them are apt to spend the last few minutes of the time allotted in marking answers to all the items they have not had time to consider. This is most undesirable because it introduces chance variance inextricably into the scores and because there is no entirely satisfactory way of correcting for chance success to prevent examinees who do this from benefiting from it.

If no penalty for marking items incorrectly is imposed, it is obvious that examinees who rush through all the items they do not have time to consider, marking almost at random, will have an advantage over those who follow directions and do not mark items they have not read. Theoretically, the best way to prevent this is to provide enough time for every pupil to consider every item. It has sometimes been suggested that no time limit be set and that pupils be allowed to leave as soon as they have tried every item. The difficulty is that, when this procedure is followed, some of the pupils will work hastily and carelessly, marking their answers almost at random in order to get away at the earliest moment. This soon becomes evident to other pupils and tends to lower the morale of the group. The confusion caused as pupils leave individually or in small groups is also troublesome; even well-intentioned pupils produce noise and distraction if they leave while others are working.

If a tryout test is administered with no time limit, all of the pupils being held until everyone has tried every item, behavior disorders are apt to occur even in well-managed schools where the morale is ordinarily good and the pupils cooperative. Experience shows that typical tryout tests of thirty to forty minutes' duration should contain enough items to permit about 90

percent of the examinees to consider every one. The time limit should be stated in advance, and no pupil should be allowed to leave before the end of the testing period except for unavoidable reasons. If a suitable time of day is chosen, this practice will ordinarily provide about as complete data as can be obtained without noticeable deterioration of pupil morale. With shorter tryout tests, the time limits can usually be set to permit more than 90 percent of the pupils to consider every item without exceeding the limits of patience for some pupils.

In these circumstances, the test constructor has to decide whether it is more unsatisfactory to use a correction for chance success or to allow the possibility that some of the examinees will disobey instructions and thereby gain an undeserved advantage. This decision is influenced by the probable appropriateness of the correction for chance success, the proportion of examinees who have probably marked items they have not seriously considered, and the extent to which these pupils will recognize that they have been able to beat the game.

Even if enough time is provided so that every examinee is able to consider every item, the proportion of items left unmarked varies from pupil to pupil. This variation is the result partly of differences in the amounts of knowledge possessed by the pupils and partly of differences in their personalities. To eliminate these personality factors, it has been proposed that pupils be required to mark an answer to every item in a tryout test. Directions to accomplish this may read, "Before handing in your test booklet and answer sheet, be sure to mark an answer to *every* question, even if you have not had time to consider it or if you feel completely unable to answer it correctly."

Even if these directions are employed, pupils will not usually follow them unless their answer sheets are inspected before acceptance and they are forced to mark an answer to every item. At one time, the writer administered a test of reading comprehension to 541 teachers-college freshmen with directions to take all the time needed and to mark every item even if this meant guessing on some items. The students worked on the items in good spirit and apparently cooperated well, but the answer sheets were not inspected before acceptance and it was later found that only 421 students had actually marked an answer to every item. To force students to mark answers to items based on reading passages available for reference is undesirable, but to force them to mark answers to items testing specific subject matter that they do not know and cannot figure out is far worse. It is not only frustrating to the students but it goes contrary to good teaching practices and compels the students to break habits of carefulness that the

schools try hard to inculcate. Once in a while it is possible that students might be told that as part of an experiment they are required to mark an answer to every item in a test even when they have no idea what to mark, but in systematic testing programs this would be inadvisable as well as impractical. It might eliminate variations in the number of omissions and thus wipe out some of the effects of differences in personality, but it would do this at a cost of antagonizing teachers and frustrating students. It would also introduce additional chance variance into the scores.

Correction for chance success is sometimes recommended partly on the ground that the percentages of success on various items will be comparable even if the items include different numbers of choices. The extent to which they actually will be comparable after correction for chance depends on the appropriateness of the basic assumption underlying correction for chance success. If the examinees answer the items *without* resorting to blind guessing or to non-chance factors uncorrelated with the function to be measured, the percentages of success will be exactly comparable even if the items include different numbers of distracters. The fact that the percentages of correct responses are usually found to increase as the number of distracters is decreased may be explained on the ground that it actually is more difficult to decide which one of several choices is the keyed answer as the number of distracters is increased; that is, for a given combination of item stem and correct response, as equally attractive distracters are added, the percentage of correct responses in comparable samples of examinees will decrease. In other words, for items of this special type, a five-choice situation is more difficult than a four-choice situation. It is doubtful whether this special type of item is encountered in actual practice any more often than the type for which misinformation and partial information play no part in determining the examinees' responses.

It should be pointed out that the corrected percentage of a sample that answers an item correctly must not be regarded as the exact percentage of the examinees that actually possesses enough information to mark the item correctly; it is, of course, only a fallible estimate of that percentage, which may be subject to systematic bias. Still less justifiably can the corrected percentage of success for an item be regarded as the percentage of a sample that "knows" the correct answer out of the context provided by the particular distracters used with it in a specific item. This is especially true of items of the best-answer type; such items are clearly intended to determine how many examinees are able to distinguish the choice keyed as correct from the particular distracters provided. The difficulty of making the distinction is a function of the number of distracters and of their

attractiveness as well as of the choice keyed as correct. In any multiple-choice item, regardless of whether it is of the best-answer type, the likelihood that an examinee will mark the correct answer depends on how successfully he can eliminate the answers keyed "wrong" as well as on how successfully he can recognize the choice keyed as "correct" or "best." This does not mean that guessing may not take place. An examinee may exclude one or more choices from consideration and then guess among those remaining; in fact, it works to his advantage to do so, whether the conventional correction for chance success is employed or not.

Summary of correcting versus not correcting for chance.—After reading the preceding discussion, the reader may well be looking for some summary that will serve as a guide to action. Let us bring together some of the main points and state their principal implications. First, it is agreed that items should be so written as to make their distracters as attractive as possible. Second, it is agreed that tryout tests should be given under as nearly unspeeded conditions as practical. Third, it is agreed that examinees should be warned not to guess blindly. Fourth, there is some disagreement regarding the extent to which blind guessing (or reliance on irrelevant non-chance factors) occurs even when it has been minimized, regarding the practicality of providing time for *every* examinee to try *every* item, and regarding the importance of preventing any examinee from thinking he has been able to beat the game.

Naturally, if a test technician believes that chance responses to a test are nonexistent, he will not use a correction for chance success. However, no one has seriously contended that such is the case; instead, some authorities believe that under attainable conditions the amount of blind guessing may be reduced to negligible proportions. Apparently these authorities are not seriously bothered by the fact that a few examinees may deliberately ignore the directions and may knowingly gain an advantage over others. Therefore, they do not recommend a correction for chance success.

Other authorities are willing to undertake the slight additional labor involved in correction for chance success even when the amount of blind guessing is of small proportions because they feel that examinees who ignore directions should realize that they may suffer a penalty for doing so. These authorities believe that the attitude of students, teachers, and laymen toward the use of tests will be better under these circumstances and that, in the long run, less unnecessary chance variance will be introduced into test scores as examinees generally come to accept the fact that no automatic premium will accompany blind guessing. It is quite true that correction cannot remove the effects of chance variance already present in

test scores, but its use can discourage the practice of blind guessing that leads to the introduction of unnecessary chance variance.

The writer has found that considerations of expense and time do not ordinarily permit tryout tests to be given in such a way that *every* pupil can try *every* item. In this situation the use of a correction for chance success tends to prevent examinees who guess blindly or mark items they have not had time to consider from profiting thereby. If most of the items are marked largely on the basis of knowledge and partial knowledge plus some guessing (as is probably often the case), the correction for chance success will tend to be too small even though the presence of misinformation will have some effect in the opposite direction.

In the writer's opinion, it is especially important to make use of a correction for chance success in obtaining individual raw scores that are to be used for internal-consistency item analysis purposes. Data obtained by Bryan, Burke, and Stewart in studies made at the Cooperative Test Service of the American Council on Education point to the conclusion that internal-consistency item analysis data for tests on which not all examinees finished are more meaningful when the high-scoring and low-scoring groups have been selected on the basis of total scores that were corrected for chance success.[3] It is probably less consequential whether

[3] Because correction for chance success often makes test scores easier to interpret, the writer is inclined to urge its use for almost all educational purposes. Recently the writer tested a group of 393 high school pupils with Form B of the Nelson-Denny Reading Test. The 47 pupils who obtained the lowest scores were then grouped together and told that they might have done better if they had marked an answer for very item whether they knew the answer or not. This is, of course, true since the Nelson-Denny test is rather highly speeded and is not scored with a correction for chance success. Form A of the test was then administered. The average score of the 47 pupils on the first test (Form B) was 25.53; on the second test (Form A) it was 46.32. According to the published norms, this difference amounted to a gain of 2.7 grades for the group. Yet, when scores on both forms were corrected for chance, the average gain was only 2 raw-score points (though the standard deviations of the distributions of scores were larger). This gain, attributable largely to regression to the population mean, is obviously a more meaningful and more readily interpreted indication of the real change that took place in the reading ability of the pupils between testings than is the gain represented by the difference in raw scores uncorrected for chance success. This is a rather unusual example but illustrates the point dramatically.

The writer discounts statements that the correlations between sets of test scores obtained with and without correction for chance are exceedingly high. In the first place, the correlations cited are usually spuriously high because they are obtained by scoring the same set of test papers in two ways. In this situation, the directions for administering the tests are the same and thus can be appropriate only to *one* of the two scoring procedures. It would be desirable to obtain the correlation between two comparable forms of the same test administered successively to the same sample with two sets of directions—one appropriate to scoring without correction for chance success and one to scoring with it. This correlation should be compared with a parallel-forms reliability coefficient for the same test based on the same sample. An exact test of the significance of the difference between the two coefficients would permit inferences to be drawn regarding the point at issue.

the percentages used in obtaining item difficulty and discrimination indices are corrected for chance success than whether the scores of the examinees on the tryout test are so corrected. There is little doubt that the selection of items for the final form of a test will be almost identical regardless of whether the item difficulty and discrimination indices have been corrected for chance success provided that proper allowance is made for the difference in level of item difficulty. However, there is reason to believe that the discrimination indices are made more nearly independent of the level of difficulty of the items when correction for chance success is employed in computing these indices (*42*, pp. 389–90). This is a point in favor of using the correction, though it is partially offset by the fact that the reliability of such indices may be a trifle lower when the correction has been made than when it has not. This change no doubt occurs because the influence of the more reliable difficulty indices is removed from the discrimination indices. The writer values the gain in meaningfulness of the discrimination indices so much more than the slight amount of reliability lost that he is inclined to favor correcting the item analysis data for chance success as well as the examinees' scores on the tryout test.

Items Omitted and Not Reached

As we have already noted, an item may not be marked for two quite different reasons: first, because the pupil reads it, realizes he does not know the answer and cannot figure it out, and deliberately omits it as he may have been instructed to do; second, because the pupil does not have time to consider the item. Unless provision is made for taking into account these two different reasons, valuable information about the items is irretrievably lost and the resulting data become difficult to interpret. To avoid this, a simple empirical check may be made. This check is based on the assumption that the items have been marked in sequence and consists in going through the answer sheets or test papers, noting which item on each sheet is the last one marked. As this is noted, a mark is made for each answer sheet on a separate tally sheet opposite the number of the next item (after the last one marked). This next item is presumed to be the first item not read by the examinee. If the criterion groups for use in the item analysis have already been selected, this check is made separately for the papers in each group. After all the papers to be used for item analysis have been examined, a cumulative frequency table is made showing the number of examinees who have not read each item; that is, have not reached it in the time limit. For the last item in a test (or in each separately timed subtest), there is no way to make a distinction between examinees who have read the item and deliberately refrained from marking an answer to it and

examinees who have not quite reached it in the time limit. If, in this dilemma, all of the examinees who did not mark an answer to the last item are considered to have lacked time to read it, the percent of correct responses computed for the item is likely to be spuriously high. On the other hand, if all of the examinees who did not mark an answer to the last item are considered to have read it and deliberately refrained from marking an answer to it, the percent computed is likely to be spuriously low. As a practical compromise, the number of examinees who did not reach the next-to-last item in the time limit may be taken as the number that did not reach the last item. This procedure provides serviceable data for the last item with the expenditure of no additional labor.

In effect, then, the percent answering an item correctly should be based on the number of examinees who marked the item plus the number of examinees who read it but decided not to mark it. In practice, this procedure prevents items that are read but not marked by quite a few examinees from appearing to be easier than they really are. It is apparent that the percent of a sample of examinees who decide to omit items depends in large measure on the directions for the test. For reasons explained previously, directions for most tests should discourage sheer guessing; hence, the examinees are apt to omit items in order to avoid a penalty for guessing incorrectly.

If, in the computation of the percent of examinees that answer an item correctly, we include the number who did not have a chance even to read the item in the time limit, we are in effect acting on the assumption that the examinees who did not reach an item would mark it correctly only as often as chance would permit. This assumption seems unreasonable to the writer, who would prefer to work on the assumption that the examinees who did not have time to read an item would have answered it correctly in about the same proportion as those who did read it.[4] This assumption seems more reasonable than any other that has practical utility, even though it is probable that the distribution of responses of examinees who did not read the item would (if they were given additional time) tend to be more like a chance distribution than like the distribution of responses made by those who did read the item (especially if the items have been arranged in order of difficulty). This expectation is based on the positive correlations usually found between proficiency (or level) scores and speed scores in any mental trait.

[4] Note that, if the criterion groups have already been identified, the procedure described in the preceding paragraphs is applied to each group separately. Thus, the assumption is that the examinees in a defined group who did not have time to reach an item in the time limit would have answered it correctly in about the same proportion as the examinees *in the defined group* who did reach it.

Formulas for Computing Percent of Correct Responses

Formulas (1), (2), and (3) are recommended for computing the percent of correct responses to an item of the types specified, provided that a correction for chance success is to be used. If this correction is not to be used, these formulas can be modified by deleting the fraction of the "wrongs" that appears in the numerator of each formula. This modification eliminates the correction for chance success but retains the adjustment for items not reached.

$$P_T = 100 \frac{R_T - \dfrac{W_T}{k_i - 1}}{N_T - NR_T}, \tag{1}$$

where P_T = the percent of correct responses in the entire sample adjusted for chance success and for omissions caused by not reaching the item in the time limit,

R_T = the number of examinees in the entire sample who answer the item correctly,

W_T = the number of examinees in the entire sample who answer the item incorrectly,

k_i = the number of choices in the item,

N_T = the number of examinees in the entire sample,

NR_T = the number of examinees in the entire sample who do not reach the item in the time limit.

If every examinee has reached every item in the time limit, NR_T becomes zero and computation of the adjusted percents is simplified accordingly. If adjustment for chance success is not required, the subtraction term in the numerator of formula (1) is omitted, again simplifying the computation.

When matching exercises are used, they should be constructed in such a way that the directions may properly specify that one of the terms in a series may be the correct answer to more than one term in the other series. To compute the percent of correct responses for each item in a matching exercise when the number of terms in each series in the exercise is the same, when more than one term in a series can be correctly matched to a single term in the other series, when there is no restriction on the order in which the examinees may mark responses to the items in each exercise, and when the same number of examinees has read every item in the exercise, the following analogue of formula (1) is recommended:

$$P_T = 100 \frac{R_T - \dfrac{W_T}{k_s - 1}}{N_T - NR_T}, \tag{2}$$

where k_s = the number of terms in each series in the matching exercise.

A matching exercise is defined as a set of related items. When formula (2) is used, it is required only that every examinee shall have marked every item in a given exercise. It is not necessary that the same number of examinees mark the items in different exercises.

A practical example to show how formula (2) is applied may be of interest. Suppose the following matching exercise has been given to 100 examinees and that the number selecting each response is as shown:

		Item 1	Item 2	Item 3
1. Boston	I. Pennsylvania	10	75	80
2. Pittsburgh	II. Massachusetts	80	15	5
3. Philadelphia	III. Connecticut	10	10	15
	Total	100	100	100

The corrected percent for item 1 in this matching exercise is found by substituting in formula (2):

$$P_T = 100 \frac{80 - \dfrac{20}{2}}{100 - 0} = 70.$$

This result indicates that approximately 70 percent of the examinees actually *know* that Boston is in Massachusetts or that it is not in Pennsylvania or Connecticut, although 80 percent of them *marked* the item correctly. For Pittsburgh the corrected percent is 62.5, and for Philadelphia it is 70. Like the correction for chance used with multiple-choice items, this procedure is a practical compromise that does not yield corrections that are strictly unbiased in each individual instance.

When a matching exercise has a different number of terms in its two series, the same basic principle is applied in correcting for chance. Again, on the assumptions that the same number of examinees have marked every item in a given exercise, that more than one term in the shorter series can be correctly matched to a single term in the longer series, and that no restriction is placed on the order in which the examinees may mark responses to the items in each exercise, the following formula is suggested:

$$P_T = 100 \frac{R_T - \dfrac{W_T}{k_l - 1}}{N_T - NR_T}, \tag{3}$$

where k_l = the number of terms in the longer series.[5]

[5] Note that formula (3) is explicitly ruled out when the terms in the longer series are to be matched to terms in the shorter series. For example, the situation in which many

The percents obtained by means of formulas (1), (2), and (3) may be used as item difficulty indices with considerable success. They can be regarded as essentially comparable regardless of the number of choices in each multiple-choice item or the number of terms in each series in a matching exercise. Needless to say, percents that are used as difficulty indices *without* correcting for chance and *without* adjusting for the failure of some examinees to read some items will be comparable only if no blind guessing occurs and if every examinee has marked every item.

A disadvantage inherent in the use of percents as difficulty indices is the fact that percents from 1 to 99 do not even approach an interval scale of difficulty. An interval scale (contrasted with a nominal or ordinal scale) possesses the properties of both order and equality in its units. It is well known that (in the case of traits that are not distributed rectangularly) a given difference between two successive percents is indicative of a greater actual difference in the underlying trait being measured as the percents being compared are moved away from 50. For this reason, when percents are used as difficulty indices, constants cannot properly be added to or subtracted from them; neither can they legitimately be averaged.

Yet, on many occasions it is convenient to express the difficulty indices of items tried out in two different samples on a single scale of difficulty or to compare the average difficulty of two sets of items tried out on the same sample. To make possible such manipulations of indices on a reasonably legitimate basis, the percents obtained by using formulas (1), (2), and (3) may be transformed into standard-deviation units. The writer has proposed that for maximum convenience to test editors this transformation be so made that the range of item difficulty indices will be 1–99 (7, pp. 7–8). If the population distribution of the variable being measured is assumed to be normal, this can be accomplished by using the Kelley-Wood Table of the Normal Probability Integral (*102,* Table 1) to transform the percents yielded by formulas (1), (2), and (3) into standard-deviation units, by multiplying each of the latter units by a constant (21.063 to five significant figures), and adding 50, algebraically, to each product. When this procedure is followed, percents of 1, 50, and 99 obtained from formulas (1), (2), and (3) will remain 1, 50, and 99, respectively, but almost every other percent so transformed will take a different value. This transformation retains the advantage of 50 as the middle of the scale and of 1 and 99 as

---

terms are to be placed in one of two categories (which comprise the shorter series) is a two-choice item and should be so corrected for chance success.

nearly extreme values while it places all intermediate indices on an approximation to an interval scale with respect to difficulty in the trait measured. For additional convenience, percents of 0 and 100 (for which the corresponding standard-deviation units are, of course, infinity) are regarded as 0 and 100, respectively, on the new scale.[6] To express the transformed difficulty indices for one set of items tried out in two different samples on a single scale with sufficient accuracy for most practical purposes, it is necessary only to average the transformed indices for each sample and to subtract the difference between the means from the transformed difficulty index of each item in the sample having the higher mean. To express the transformed indices for two different sets of items tried out on different samples, an identical group of representative items may be inserted in each set. The difference between the averages of the transformed indices for this group of items in each sample may then be obtained and used as a constant for subtraction from the transformed index for each item in the set accompanied by the higher mean for the group of identical items. To compare the average difficulty of two different sets of items tried out on the same sample, the transformed indices may be averaged and the difference noted.

Experimental evidence has shown that difficulty indices of the sort described are extremely reliable when they are based on samples as large as 400. The convenience of using the same percents for obtaining both difficulty indices and discrimination indices has led to the computation of difficulty indices that are only *estimates* of the percents in the entire sample. These estimates are sometimes based on data obtained from only the highest and lowest 27 percent of the sample tested and rest on the assumption that the regression of individual item scores on the scores used for selection of the highest and lowest 27 percent of the sample is rectilinear. This assumption is not unreasonable in most cases and makes possible considerable economy of labor. It should, however, be subjected to experimental verification.

The writer has computed the reliability coefficient of a group of typical item difficulty indices estimated in this way and has found it to be .98 when the sample included 100 examinees in the highest 27 percent and 100 examinees in the lowest 27 percent (*42*, pp. 385–90; *41; 47*). These data suggest that the loss of reliability incurred by estimating indices from only 54 percent of the sample tested is not sufficient to be of practical consequence when the two criterion groups employed include at least 100 examinees apiece.

[6] A table for transforming percents into these indices is provided by Davis (7, Table 4).

In actual practice, then, formula (1) may be replaced by the following:

$$P_{\text{Est}} = 100\,\frac{\dfrac{R_H - \dfrac{W_H}{k_i - 1}}{N_H - NR_H} + \dfrac{R_L - \dfrac{W_L}{k_i - 1}}{N_L - NR_L}}{2}, \tag{4}$$

where P_{Est} = the estimated percent of correct responses in the entire sample adjusted for chance success and for omissions caused by not reaching the item in the time limit,

R_H = the number of examinees in the highest 27 percent of the sample who mark the item correctly,

R_L = the number of examinees in the lowest 27 percent of the sample who mark the item correctly,

W_H = the number of examinees in the highest 27 percent of the sample who mark the item incorrectly,

W_L = the number of examinees in the lowest 27 percent of the sample who mark the item incorrectly,

N_H = the number of examinees in the highest 27 percent of the sample,

N_L = the number of examinees in the lowest 27 percent of the sample,

NR_H = the number of examinees in the highest 27 percent of the sample who do not reach the item in the time limit,

NR_L = the number of examinees in the lowest 27 percent of the sample who do not reach the item in the time limit,

k_i = the number of choices in the item.

Formula (2) becomes:

$$P_{\text{Est}} = 100\,\frac{\dfrac{R_H - \dfrac{W_H}{k_s - 1}}{N_H - NR_H} + \dfrac{R_L - \dfrac{W_L}{k_s - 1}}{N_L - NR_L}}{2}, \tag{5}$$

and formula (3) becomes:

$$P_{\text{Est}} = 100\,\frac{\dfrac{R_H - \dfrac{W_H}{k_l - 1}}{N_H - NR_H} + \dfrac{R_L - \dfrac{W_L}{k_l - 1}}{N_L - NR_L}}{2}. \tag{6}$$

Formulas (4), (5), and (6) may be modified if correction for chance success is not to be used simply by deleting the fraction of the "wrongs"

in the high and low groups indicated for subtraction. This modification retains the adjustment for items not reached.

The proposal has sometimes been made that item difficulty indices be transformed into scores having widely known characteristics. For example, the Cooperative Achievement Tests ordinarily yield Scaled Scores that possess properties with which many test users and test constructors are familiar. It would be possible to equate difficulty indices for the items in an experimental test to these Scaled Scores. This could be done most easily if the item were assumed to measure the same trait as one of the Cooperative tests for which Scaled Scores have been provided, but it could be accomplished whenever the correlation of scores (pass or fail) on the item with scores on such a test was known. Although some advantages would accrue if this proposal were carried out, they might not be great enough to warrant the expenditure of time and labor required.

Before discussing the use of difficulty indices in item selection, we shall consider the measurement of item discriminating power.

Item Discriminating Power

For many years test constructors have recognized the need for determining the value of each test item for making the test in which it is included rank examinees accurately with respect to a defined criterion variable. If the test constructor knew the relative value of each item, he could select only the best for inclusion in the final form of the test. Unfortunately, there is no analytic solution to this problem that is even reasonably practical. Theoretically, if suitable criterion scores were available, product-moment correlation coefficients between them and the scores on each individual item could be computed. Likewise, the intercorrelations of the items could be obtained. Then a multiple regression weight for each item could be obtained and items having weights significantly different from zero at a specified level of confidence could be retained in the final form of the test. A requirement inherent in this procedure, which renders it of theoretical interest only, is that the variances of all the principal components of the matrix of item intercorrelations would have to be proved significantly different from zero at an appropriate level of confidence. Obviously, this could not be done without using samples of fantastic size. Since regression coefficients for individual items that are obtained on the basis of samples of ordinary size (say, up to 1,000 cases) are not meaningful, the approximation procedures that may be used by test constructors are probably more defensible as well as more practical. Later in this chapter we shall discuss these procedures. First, it may be best to consider specific methods of expressing the item-criterion relationship.

The Criterion Variable

Whenever it is practical, test constructors should obtain as nearly unbiased measures as possible of the ultimate criterion variable that is appropriate for the test being built. That these measures be highly reliable is much less important than that they be unbiased. For most achievement tests, however, it is difficult or impossible to obtain criterion measures. As pointed out in chapter 6 on "Planning the Objective Test," the content of an achievement test is often formulated by analyses of curriculums and textbooks and by the pooled judgment of recognized authorities in the field. Under these circumstances a well-constructed test may constitute the best available measure of the criterion; in a sense, the test itself defines the function it is to measure. Such tests may be described as self-defining.

The total scores derived from a self-defining test are, therefore, often used as the immediate criterion measures with which the individual items in the test are correlated. These relationships between the total scores derived from a test and item scores are referred to as internal-consistency item discrimination indices. The relationships between item scores and scores in a criterion variable other than the total score on the test are referred to as "item validity indices." The term "item discrimination indices," as used in this chapter, includes both internal-consistency item discrimination indices and item validity indices.

When the total score on a test is used as the criterion variable for judging the discriminating power of each item in it, the resulting indices reflect, among other influences, the extent to which the item measures the same mental functions as the total score. The fact that some items prove to have more discriminating ability than others means that for the group tested they are better measures of whatever the whole test actually measures. Except when all of the items in a test measure the same mental functions with less than perfect reliability, internal-consistency item analysis data provide an inadequate basis for comparing the discriminating power of items for measuring whatever each individual item actually does measure. If two items of the same difficulty that have identical discriminating power for measuring two different mental traits are tried out with 98 homogeneous items, the one of the two that measures a trait more heavily weighted in the 98 items than the other will turn out to have the higher discrimination index. This is an obvious point but easily overlooked. It is an important point, too, because in practice we try out items in rather heterogeneous groups. Suppose, for example, that in a group of reading items of similar difficulty 90 call for finding detailed facts while 10 call for making inferences. If discrimination indices are computed for these

items, using the total score on the 100 reading items as the criterion variable, the 10 inference items will be found to have markedly lower discrimination indices than the items that call for finding detailed facts. If a 50-item reading test is to be constructed by selecting the items having the highest discrimination indices, no inference items are likely to be included. Yet experts in the field of reading would criticize the validity of such a test as a measure of reading ability. To avoid situations like this, it is desirable to assemble items into closely homogeneous groups for item analysis purposes. For example, items that call for finding detailed facts and for making inferences should be grouped separately. This may sometimes be impracticable, as would be the case if only 10 items of a given type were available.

Methods of Expressing Item Discriminating Power

Many ingenious statistical procedures have been devised to provide discrimination indices. Some of these are so simple and require so little labor that teachers may profitably make use of them for ordinary classroom tests. Suppose, for example, that fifty pupils mark answers to each of fifty true-false items on material studied during the preceding week. Usually classroom tests of this kind are given with enough time for every pupil to try every item. The teacher may score the tests and put the test sheets in order by size of score. From the pile of papers he may take the top quarter (twelve to thirteen papers) and the bottom quarter and simply tally the number of correct answers to item 1 in each quarter. After this has been done for all fifty items, he can discard items with approximately the same number of tally marks in the two groups or with more tally marks in the bottom than in the top quarter. This procedure will tend to identify items with little internal-consistency discriminating power and lead to greater efficiency of measurement in a revision of the test composed of the more discriminating items. The fact that the procedure is crude and subject to a variety of errors should not deter teachers from making use of it. On the other hand, for professional work in test construction it is not adequate.

A variation of well-known graphic methods of item analysis has been described by Turnbull (*32*). After dividing the tryout group into sixths on the basis of some criterion (such as total test score), the test constructor plots on specially prepared graph paper the percent of the examinees in each sixth who marked the correct answer to each item. The percent who marked each incorrect choice is also plotted. Each sheet of graph paper then indicates the discrimination ability of the correct answer and each of the incorrect choices for a given item. This process provides revealing data at

the cost of a great deal of labor. Short cuts are provided, however, and a means of estimating an item-criterion correlation coefficient is outlined. The method might be useful if large numbers were tested and if IBM equipment were available.

Critical ratios have sometimes been suggested as measures of item discriminating power: the percents of correct responses in two separate criterion groups are obtained, and the differences between these pairs of percents are computed for comparison with their own standard errors. The resulting critical ratio for each item does indicate *how likely* it is that a given item actually differentiates between the two criterion groups, but it does not usually make conveniently possible a comparison of the *amount* of discriminating power possessed by a number of items. It fails in this respect because its magnitude is in part dependent on the number of individuals in the two criterion groups who marked a response to each item. In practice, this number may sometimes decrease from item to item in each group because some examinees ordinarily fail to reach items toward the end of a tryout test.

Chi square has been proposed as a measure of item discrimination by Guilford, and others (*16*). In an unpublished study, Cureton has suggested using a variation of the chi-square test for the same purpose. But, as the latter has pointed out, critical-ratio and chi-square tests are applicable only to large samples and to item choices that are in the middle of the range of attractiveness to the examinees. To circumvent this limitation, Cureton has proposed the use of a chi test. However, all of these tests suffer from the limitation mentioned in connection with the use of the critical ratio as a measure of item discrimination; namely, that they do not usually make conveniently available a comparison of the *amount* of discriminating power of a group of items. On the other hand, indices that permit easy comparison of the amount of discriminating power of a group of items do not lend themselves to exact tests of significance.

The chi test proposed for use by Cureton is designed to determine, at designated levels of confidence, whether a sample drawn at random from the population in which the correct answer to the item is marked by equal proportions of the two criterion groups will have the proportion of the high-scoring criterion group marking the correct answer as large as or larger than the proportion of the low-scoring criterion group marking the correct answer. Inasmuch as this test is not widely known and is applicable to a greater degree than most others when the samples are small, it is outlined briefly here. The chi test involves no assumptions regarding the shape of the distribution of the traits measured by either the criterion or item scores.

Chi is computed as follows if the number of examinees in the high-scoring group who mark the item correctly is greater than the number in the low-scoring group who mark it correctly:

$$Chi = \frac{R_H - R_L - 1}{\sqrt{R_T\left(1-\frac{R_T}{N_T - NR_T}\right)}}. \tag{7}$$

Chi is computed as follows if the number of examinees in the high-scoring group who mark the item correctly is smaller than the number in the low-scoring group who mark it correctly:

$$Chi = \frac{R_H - R_L + 1}{\sqrt{R_T\left(1-\frac{R_T}{N_T - NR_T}\right)}} \tag{8}$$

where R_H = the number of examinees in the high-scoring group who mark the item correctly,

R_L = the number of examinees in the low-scoring group who mark the item correctly,

$R_T = R_H + R_L$,

N_T = the number of examinees in the high-scoring and low-scoring groups,

NR_T = the number of examinees in the high-scoring and low-scoring groups who do not reach the item in the time limit.

For reasons to be explained later in this section, the high-scoring and low-scoring groups should each constitute about 27 percent of a group for which continuous criterion scores are available. In any case, the two groups should be of equal size to make use of the significance levels shown in Table 6. This table is an abridgement made by Cureton of Table 8 in *Statistical Tables for Biological, Agricultural, and Medical Research* (*100*).

Among statistics proposed for expressing item discriminating power in the form of relationships, the biserial product-moment r (sometimes called the point biserial r), the biserial r, the tetrachoric r, and the phi coefficient have been suggested. The choice among these depends partly on the purpose for which the test and the item analysis data are to be used, partly on the convenience with which each statistic serves that purpose, and partly on the ease and economy of computation required by the practical circumstances.

If the test constructor is primarily interested in constructing a predictor test for use with a group of examinees having essentially the same level and distribution of ability as the group used for item analysis purposes, the biserial product-moment r may be used when the criterion is a continuous

TABLE 6

VALUES OF CHI AT VARIOUS SIGNIFICANCE LEVELS FOR CERTAIN SAMPLE SIZES

VALUE OF $\frac{R_T}{2}$ IF FORMULA (7) IS USED; *or* VALUE OF $\frac{N_T - NR_T - R_H + R_L}{2}$ IF FORMULA (8) IS USED	VALUE OF CHI AT SIGNIFICANCE LEVEL	
	.05[a] *or* .025[b]	.01[a] *or* .005[b]
2	1.90	......
3	1.91	2.44
4	1.92	2.46
5	1.93	2.48
6	1.93	2.50
7–10	1.94	2.52
11–18	1.95	2.54
19–36	1.95	2.56
37–100	1.96	2.57
101 and over	1.97	2.58

[a] If formula (8) is used.
[b] If formula (7) is used.

variable and will be used as such. This statistic would be especially desirable in the rather unusual case when the item-criterion relationships were to be used to obtain multiple regression weights for each item. Formulas for computing biserial product-moment coefficients and their variance errors are presented by Kelley (*101*, pp. 370–73). A convenient computing formula is as follows:

$$r_{\text{pbis}} = \frac{M_R - M_{(N_T - NR_T)}}{\sigma_{(N_T - NR_T)}} \sqrt{\frac{p_R}{1 - p_R}}, \tag{9}$$

where M_R = the mean criterion score of examinees who mark the item correctly,

$M_{(N_T - NR_T)}$ = the mean criterion score of examinees who reach the item in the time limit,

$\sigma_{(N_T - R_T)}$ = the standard deviation of criterion scores of examinees who reach the item in the time limit,

p_R = the number of examinees who mark the item correctly divided by the number who reach the item in the time limit.

If the criterion variable is a natural dichotomy and must be used as such, an acceptable method of expressing the item-criterion relationship is the phi coefficient when the group with which the test is to be used is essentially the same with respect to level and distribution of ability as the group used for item analysis purposes and when the point of dichotomy in the criterion variable remains constant in successive groups in which

the test is being used for prediction purposes. When these extremely unusual conditions are met, the phi coefficient is appropriate because it is a rigorous product-moment *r* subject to precise tests of significance and suitable for use in computing multiple regression weights. Computational routines for obtaining phi coefficients and their variance errors are illustrated by Kelley (*101*, pp. 379-82).

A table from which phi coefficients may be read directly has been prepared by Jurgensen (*53*) for use in the special case when the number of examinees in each of the dichotomous criterion groups is the same. This condition is rarely met precisely in item analysis work because some examinees usually fail to reach certain items in the time limit, but Jurgensen's table might be of considerable value when every examinee (or nearly every one) has marked an answer to every item.

An abac published by Guilford (*16*) makes possible estimation of the phi coefficient from percents of correct responses made by examinees in high-scoring and low-scoring groups of a distribution of criterion scores. Unless these groups consist of the highest and lowest 50 percent of the sample, this procedure leads to coefficients that take values having no analytical relationship to the product-moment *r* and that are not subject to precise tests of significance. Thus, they do not possess the two properties that justify the use of the phi coefficient under the circumstances described previously, but they are economical to compute.

To avoid some of the disadvantages of the phi coefficient and the biserial product-moment coefficient, the biserial and tetrachoric correlation coefficients have been suggested as indices of item discriminating power. Both of these require the assumption that a normally distributed underlying variable has been forced into a dichotomy. Computation of the biserial *r* demands this assumption only in the case of one variable, while computation of the tetrachoric *r* requires acceptance of this assumption for both variables. It is clear that when an assumption of normality is made without evidence to support it, any tests of the significance of coefficients computed on the basis of it may be misleading. Only when cogent reasons for doing so are presented should the test constructor abandon product-moment coefficients for estimates of relationships expected in parent distributions assumed to be normally distributed.

As it happens, such reasons often present themselves. It is sometimes difficult or even impossible to obtain for experimental purposes samples of examinees in which the level and distribution of ability are reasonably similar to those in the groups with which the final form of the test is to be used. Sometimes a test is to be used in several school grades. In this case

the selection of items for the final form should theoretically take into account separately the discriminating power of each item among pupils in each grade in which the test is to be used. Given this situation, the problem of item selection can become very complicated, and test editors often seek to simplify it by using for each item one discrimination index that is unaffected by the average level of ability of the pupils and that is based on a weighted proportion of pupils at each of several grade levels.

A different kind of consideration that often leads the test constructor to prefer an index of item discriminating ability that is essentially unaffected by the level of difficulty of each item is the efficiency of measurement that may sometimes be obtained by deliberate control of the distribution of item difficulty levels when a test is assembled. It is a well-known fact that the variance of any single item is at a maximum when it is used in a group of examinees in which 50 percent mark it correctly. From this it logically follows, as Gulliksen has demonstrated, that test-score reliability and variance are maximized when every item is of 50 percent difficulty regardless of the level of item intercorrelation (*75*). Unfortunately, many test constructors have uncritically accepted high over-all reliability as the primary goal in test construction. Sometimes it should be so accepted, but often the over-all reliability coefficient of a test is merely an interesting but irrelevant statistic. Ordinarily, the primary goal in test construction is to maximize the number of discriminations among all the examinees or between such groups of examinees as the test administrator designates. Methods of selecting items to approximate this goal will be discussed in some detail in a later section of this chapter.

To provide an index of discriminating ability that is essentially unaffected by differences in the percent of testees answering correctly items scored "right" or "wrong," the biserial r may be employed when the criterion variable is continuous. One of the computing formulas for this statistic, which is given by Dunlap (*10*), may be written in the notation of this chapter as follows:

$$r_{\text{bis}} = \frac{M_R - M_{(N_T - NR_T)}}{\sigma_{(N_T - NR_T)}} \cdot \frac{p_R}{z}, \tag{10}$$

where $z =$ the ordinate in the unit normal distribution which divides the area under the curve into the proportions p and q.

If the criterion scores of the examinees have been corrected for chance by means of an appropriate formula, the procedure described above will provide correlation coefficients based on corrected scores and thus com-

parable in this respect to other types of discrimination indices to be described later in this chapter.

The variance error of biserial r may be approximated when an item is marked correctly by from 5 percent to 95 percent of the testees who read it by the following formula (which assumes that the distribution underlying the dichotomy is exactly normal and that the sample is large):

$$V_{r_{\text{bis}}} = \frac{\left(\frac{\sqrt{pq}}{z} - r_{\text{bis}}^2\right)^2}{N_T - NR_T}, \tag{11}$$

where $q = 1 - p$.

It is always desirable to make use of item analysis data for which variance errors may be computed, yet, in practice, tests of significance may often be of little utility because the test constructor ordinarily has to make use of the best items he has even if the relationships of some of them with the criterion variable are not significantly different from zero. When a test of significance can be utilized, however, the variance error of zero r_{bis} is often found most useful. Formula (11) may be simplified to compute the statistic:

$$V_{\text{zero } r_{\text{bis}}} = \frac{pq}{z^2(N_T - NR_T)}. \tag{12}$$

If the criterion to be used for item analysis purposes is a natural dichotomy, we cannot use biserial r as an index of item validity for items scored only "right" or "wrong." A situation of this kind arises when we wish to correlate scores on each of a set of items with a dichotomous criterion variable such as "graduated from college with a degree" and "did not graduate from college with a degree," or "completed pilot training and received his wings" and "did not complete pilot training and receive his wings." In this situation, if we want to obtain an index of item validity that is comparable to the biserial r, the tetrachoric correlation coefficient should be employed. Its computation assumes that the distributions underlying the two dichotomies are exactly normal. Needless to say, these assumptions should never be made unless there is good reason for believing them to be true, or so nearly true that the data obtained by making them will not lose their serviceability.

The computation of tetrachoric coefficients by formula is so laborious that it is never attempted for practical purposes when a large number of coefficients is required. Tables to facilitate the computation of tetrachoric

r have been available for many years (*105,* Tables 29-30) and, more recently, computing diagrams have become available for the same purpose (*98; 51*). The writer recommends these computing diagrams when tetrachoric r is to be used as a measure of item validity. Goheen and Kavruck have recently published a work sheet that may be found convenient for use with Hayes' diagrams (*15*). The computational routine for use with the diagrams prepared by Chesire, Saffir, and Thurstone is presented here in order to show how the correction for chance and the adjustment for examinees who did not reach some items in the time limit may be incorporated in the procedure. A fourfold table is first set up with each of the cells lettered as in the accompanying diagram.

A	B	C
D	E	F
G	H	1.000

Entries in the cells of this table may be filled in by means of the following formulas:

$$B = \frac{R_H - \dfrac{W_H}{k_i - 1}}{N_T - NR_H - NR_L}, \tag{13}$$

$$C = \frac{N_H - NR_H}{N_T - NR_H - NR_L}, \tag{14}$$

$$E = \frac{R_L - \dfrac{W_L}{k_i - 1}}{N_T - NR_H - NR_L}, \tag{15}$$

$$A = C - B, \tag{16}$$

$$D = F - E, \tag{17}$$

$$F = 1 - C, \tag{18}$$

$$G = A + D, \tag{19}$$

$$H = B + E, \tag{20}$$

where R_H = the number of examinees in the high criterion group who mark the item correctly,

R_L = the number of examinees in the low criterion group who mark the item correctly,

W_H = the number of examinees in the high criterion group who mark the item incorrectly,

W_L = the number of examinees in the low criterion group who mark the item incorrectly.

N_T = the number of examinees in both criterion groups,

N_H = the number of examinees in the high criterion group,

NR_H = the number of examinees in the high criterion group who do not reach the item in the time limit,

NR_L = the number of examinees in the low criterion group who do not reach the item in the time limit,

k_i = the number of choices in the item.

The appropriate cells in each fourfold table are used to enter the computing diagrams, and the tetrachoric r is obtained with a minimum of labor. The entry in cell H may be multiplied by 100 (to convert it into a percent) and used as an index of item difficulty, for it will be comparable to the adjusted percent provided by formula (1).[7]

The variance error of a tetrachoric correlation coefficient can only be approximated, but there are occasions on which the test constructor is willing to utilize even a dubious variance error to set some sort of objective standard for rejecting items that do not appear to possess significant relationships with the criterion (at a specified level of confidence). The following formula may be employed to estimate the variance error of a tetrachoric r when the true value is zero:

$$V_{r_{\text{tet}}} = \frac{pqp'q'}{z^2 z'^2 (N_T - NR_T)}, \tag{21}$$

where p = the proportion of those examinees that reach the item in the time limit who are in the high criterion group,

$$q = 1 - p,$$

$$p' = \frac{R_T - \dfrac{W_T}{k_i - 1}}{N_T - NR_T},$$

$$q' = 1 - p',$$

z = the ordinate in the unit normal distribution which divides the area under the curve into the proportions p and q,

z' = the ordinate in the unit normal distribution which divides the area under the curve into the proportions p' and q',

[7] If desired, these adjusted percents can be immediately transformed into corresponding values on the scale of difficulty indices proposed by Davis (7, Table 4).

N_T = the number of examinees in the sample,

NR_T = the number of examinees in the sample who do not reach the item in the time limit,

R_T = the number of examinees in the sample who mark the item correctly,

W_T = the number of examinees in the sample who mark the item incorrectly,

k_i = the number of choices in the item.

An example of a situation in which the use of tetrachoric r as an index of item validity is to be preferred is provided by the validation of items constructed for use in the Aviation Cadet Qualifying Examination against the criterion of passing or failing in pilot training in the Army Air Forces. This criterion is a good example of a variable in which the underlying ability (to learn to fly an airplane) is probably normally distributed but is expressed as a dichotomy. Yet the proportion of cadets who were assigned scores of "pass" varied markedly because of administrative considerations entirely irrelevant to the psychological factors involved. Scores on each multiple-choice item used in the Aviation Cadet Qualifying Examination were expressed only as "right" and "wrong," though it is reasonable to suppose that a normal distribution of talent underlay the responses to each item, especially to items not extremely easy or difficult. For practical reasons, the items had to be tried out in groups of aviation students already selected by previous forms of the test. These students constituted a far more able and more homogeneous group than the applicants for cadet training with whom the final form of the test was to be used. In this situation, the tetrachoric r obtained on the basis of the selected group is clearly a better estimate of the correlation of the underlying variables in a sample of applicants than the phi coefficient or any rigorous product-moment r.

A method for estimating tetrachoric correlation coefficients for use as indices of item discriminating power that is somewhat more economical than the method described above for use with the computing diagrams was suggested in 1940 (*28*). This method is not applicable except when either the upper and lower 50 percent or the upper and lower 25 percent of the testees with respect to the criterion variable can be identified. In other words, the method cannot ordinarily be used if the criterion variable is a natural dichotomy.

Vernon has suggested the use of what he calls double tetrachoric coefficients. He has shown that these are more reliable than ordinary tetrachoric coefficients, yet may be obtained very rapidly (*33*).[8]

[8] This excellent article by Vernon presents a summary of item analysis techniques with some evaluative commentary.

From the preceding discussion in this chapter, it is apparent that the type of item analysis data to be selected for use in constructing a test is dependent on the purpose of the test constructor and the kind of basic data he can obtain. There is no one type of item analysis data that is best under all circumstances. Nevertheless, circumstances that favor the use of biserial *r* as an index of discriminating power appear to be most numerous. These circumstances include the fact that the level of distribution of ability is not always the same in experimental and consumer groups of examinees, and the fact that it is rarely desirable for a test to have its minimum standard error of measurement exactly at the raw-score point corresponding to the average level of ability in the group tested.

Unquestionably, biserial correlation coefficients would have been used more frequently in the past for item analysis purposes if they were not so laborious and expensive to obtain. To meet the need for an economical approximation to biserial coefficients, Flanagan, working with Kelley, devised an ingenious procedure based on the fact that, since the magnitude of a correlation coefficient is determined by extreme cases to a much greater extent than by cases near the middle of the bivariate surface, an estimate of the coefficient may be obtained with a much greater decrease in labor than in accuracy by utilizing only the data in the tails of the two distributions. This procedure became of practical utility for estimating correlation coefficients in 1931 when Part II of Pearson's *Tables for Statisticians and Biometricians* (*106*) was published. Tables 8 and 9 provide the frequencies on a normal bivariate surface for cells one-tenth of a standard deviation square having lower limits of 0.0, ±0.1, ±0.2, . . ., ±2.5 standard deviations from either or both means. All cells having lower limits of ±2.6 standard deviations extend to infinity. These frequencies are provided for product-moment correlation coefficients at intervals of .05 from −1.00 to +1.00.

Given these data, all that remained to make the procedure practical was to determine the best proportion of the tails of the criterion distribution to be employed and to construct tables for convenient use. Kelley demonstrated in 1939 (*19*) that for items of 50 percent difficulty and low reliability scored in graduated amounts the optimum proportions are 27 percent in each tail of the criterion distribution. Though items are not always close to the 50 percent level of difficulty and are rarely scored in other than two categories ("right" and "wrong"), Kelley concluded that the upper and lower 27 percent of the criterion distribution are ordinarily most serviceable.

Empirical evidence to support this conclusion was obtained prior to 1942 in unpublished studies at the Cooperative Test Service; in a report on

Printed Classification Tests, edited by Guilford and Lacey (*49,* pp. 30–31), some evidence of this kind has been published. Internal-consistency item discrimination indices were computed by several methods for 68 items in a test called Visualization of Maneuvers. Two separate samples of 400 aviation cadets were employed. The product-moment correlations between indices based on these two samples were obtained to provide a measure of the reliability of indices computed by the various methods employed. The reliability coefficient for biserial *r*'s is .87, for estimates of the biserial *r* obtained from Flanagan's table it is .87, and for tetrachoric *r*'s it is .79.

The use of Kelley's method assumes that normal distributions of talent underlie the dichotomous item response categories of "right" and "wrong" and the criterion distribution of which only the highest and lowest 27 percent are utilized. Rectilinearity of regression of item scores on criterion score is also assumed. It is evident, therefore, that the correlation coefficients estimated by means of Flanagan's table are strictly analogous to tetrachoric correlation coefficients, though more reliably determined than the latter.

In 1935 Flanagan published an abbreviated table of product-moment correlation coefficients (*45,* Table 11) and in 1936 made available *A Table of the Values of the Product-Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of Successes.* Later, he presented a valuable discussion (*12*) of the use of coefficients obtained by means of this table, which has been widely used to obtain item-criterion correlation coefficients. Moreover, Flanagan's table has general utility and may be found exceptionally useful in many instances outside the field of test construction where economical approximations to biserial correlations are required.

The variance errors for coefficients of correlation read from Flanagan's table cannot now be determined by analytical means, but it is evident on the basis of both theoretical considerations and empirical data that their sampling errors are larger than those of biserial *r*'s and smaller than those of tetrachoric *r*'s. This may come as a surprise to many research workers who may have supposed that the reliability of the resulting data would be impaired by eliminating the middle 46 percent of the cases. But Kelley suggested that the elimination of this group should actually improve the reliability over what it would be for tetrachoric coefficients computed on the basis of the highest and lowest 50 percent of the same sample, and empirical evidence, such as that reported above, confirms this expectation.

Data concerning the reliability of item-criterion correlation coefficients

obtained by means of the Flanagan table have been published (*43; 44*). For 86 items pertaining to information about flying, the writer computed two sets of item-test correlation coefficients, using two comparable samples of 370 aviation students. Each student tried every item. The high-scoring and low-scoring groups were chosen on the basis of the total score derived from 81 of the 86 items. The correlation of the two sets of item-test correlation coefficients proved to be .67.[9] The standard error of measurement in this particular group of item discrimination indices was approximately .08.

In those rare circumstances when the use of the product-moment r, the product-moment biserial r, or the phi coefficient as measures of discriminating ability is appropriate, the obtained coefficients have certain inherent properties of great value.[10] Consequently, the test constructor has no desire to express them otherwise. However, if the data take the form of biserial r's, estimates of biserial r's from Flanagan's table, or tetrachoric r's, the test constructor may regard them as indices of item discriminating ability essentially unaffected by differences in the level of difficulty of the items or in the variability of criterion scores for the group that reaches successive items. He may then wish to transform them into more nearly an interval scale of discriminating power.

It is well known that it becomes more and more difficult to raise a correlation coefficient by a certain number of hundredths as perfect correlation is approached. A difference of .05 between correlation coefficients of .90 and .95 represents a far greater difference in relationship than a difference of .05 between coefficients of .05 and .10. This is not a serious limitation for item analysis purposes, but in the case of product-moment coefficients it is so easily removed that there is no reason to tolerate it if it works any inconvenience. To satisfy the special requirements of test editors for an index of discriminating power that is substantially comparable from item to item and that is easy to use, the writer has suggested indices that constitute a linear function of Fisher's z and range from 0–99, thus eliminating decimals. They have properties that permit them legitimately to be added, subtracted, and averaged; their variance errors are virtually identical regardless of their magnitudes when they are based on samples of the same size or essentially the same size; and the units in which they are expressed are sufficiently coarse to discourage an impression of ex-

[9] Mean of set A = .29; SD of set A = .15; mean of set B = .28; SD of set B = .13; N = 86.

[10] This is not true of the phi coefficient if it has been obtained from an abac such as Guilford's, unless the high-scoring and low-scoring groups each consist of 50 percent of the sample.

treme precision of measurement, yet fine enough to satisfy all practical requirements of test construction.

To minimize the labor required for obtaining item analysis data, a chart has been prepared which, like the Flanagan table, is entered with percents of correct responses obtained by examinees in the lowest 27 percent and the highest 27 percent of the criterion-score distribution. Unlike the Flanagan table, however, this chart yields for most usable test items both an index of discriminating power and an index of difficulty in one operation.[11] These indices possess the properties recommended previously in this chapter. Their variance errors, like those of the correlation coefficients derived from Flanagan's table, cannot now be obtained analytically, but it is known that they are more reliably determined than tetrachoric r's and less reliably determined than biserial r's. Empirical evidence concerning the correlation between two groups of item discrimination indices obtained from the Davis item analysis chart was presented by the writer in 1946 (*42*). For 86 items the correlation coefficient based on two comparable samples of 370 testees was .58.[12] When the percents used to enter the chart were *not* corrected for chance, the corresponding correlation coefficient became .65.[13]

The correlation of the discrimination indices with difficulty indices for the same items was .11 in sample A when the percents used to enter the chart were corrected for chance. When the correction was not made, the correlation of the resulting sets of indices in sample A was .41. These data show that correction for chance greatly reduced the correlation between these discrimination and difficulty indices and suggest that when the Davis item analysis chart is used in the way recommended, it yields discrimination and difficulty indices that are essentially independent.

Unquestionably, improved methods of item analysis will be developed. Application of the principles of sequential sampling may provide one of these. In fact, Walker has found that this technique does efficiently separate, at a designated level of confidence, items that have a significant relationship with the criterion from those that do not (*36; 37*). Information comparable to that supplied by the more familiar critical-ratio, chi-square, or chi tests is provided, but the sequential-sampling method is said greatly to reduce the labor involved if specially prepared tables are used. Some of the latter have already been made available.

[11] Computational procedures for use with this chart have been provided in great detail. (7, chap. 5).

[12] Mean (sample A) = 25.14; mean (sample B) = 24.10; SD (sample A) = 14.30; SD (sample B) = 11.75; N = 86.

[13] Mean (sample A) = 18.73; mean (sample B) = 18.03; SD (sample A) = 10.14; SD (sample B) = 7.85; N = 86.

The use of factorial methods for item analysis purposes awaits the development of electronic computers capable of extracting factors from large matrices of intercorrelations with great rapidity. As soon as these intruments become widely available, it may be that items will be selected on the basis of their factor loadings in the principal components of the matrix of intercorrelations obtained by intercorrelating all the items in a test. Many technical and practical problems will have to be solved, however, before factorial methods will yield meaningful data for item analysis purposes. If product-moment correlation coefficients are used in the matrix to be analyzed, the varying difficulties of the items will create trouble, as indicated by Lawley (*80*). If tetrachoric coefficients are used, their unreliability and lack of susceptibility to precise tests of significance will raise awkward problems. In any event, we can look forward with interest to developments in the field of item analysis that even now we can see taking shape.

Factors Affecting the Interpretation of Item Discrimination Indices

Spurious Correlation in Internal-Consistency Data

When the total score on a test is used as the criterion variable for evaluating each individual item in the test, the critical ratios, chi squares, chi's, or item-test correlation coefficients that are computed are spuriously high. This results from the fact that each item is part of the criterion with which it is correlated or compared. One way of eliminating this overlapping is to score the test as many times as there are items in it and to use as the criterion for evaluating each item a total score from which it has been excluded. However, this procedure is impractical because it requires an expenditure of labor out of all proportion to the benefit derived from it. Unfortunately, there is no statistical technique by which the effect of the overlapping can be accurately removed with an increase in computational labor small enough to justify the resulting benefit.[14] The best that can be done is to indicate what the order of magnitude of the spurious correlation is likely to be and point out that the relative magnitudes of the item discrimination indices are affected less than their absolute magnitudes.

The smaller the number of items in the total score used as a criterion variable, the larger will be the spurious item-criterion relationship. Consequently, items should be tried out in large groups when internal-consistency item analysis data are to be obtained. It is evident that the length of the tryout test is usually an important consideration in the interpretation of internal-consistency item discrimination indices.

[14] To remove the spurious element from biserial coefficients, *see 38,* equation 4.

For discrimination indices expressed as correlation coefficients, the magnitude of the spurious element can be calculated precisely for two limiting conditions (7). If a set of items are all of the same level of difficulty (whatever that may be), it can be shown that when the intercorrelations are all zero, the spurious item-criterion correlation coefficient will be $\frac{1}{\sqrt{n}}$, where n equals the number of items. For a 100-item test, therefore, the spurious correlation will be .10. If the item intercorrelations are all unity, the item-criterion correlation will necessarily be unity and no spurious element will be present. As this illustration suggests, the spurious element in the item-criterion coefficient decreases as the average item intercorrelation increases. Fortunately, the size of the spurious element drops rapidly with the first few hundredths of item intercorrelation. One other element that is important in determining the amount of spurious correlation is the difficulty of the item. The spurious element is maximized when an item is of 50 percent difficulty; the discrimination indices for easy and hard items need not be discounted so much as those for items of median difficulty.

Position of an Item in the Tryout Form

Generous time limits should always be provided so that all, or almost all, of the examinees will have a chance to try every item. However, it is sometimes impossible to arrange administrative conditions to make this readily possible. Then, items near the end of the tryout test may not be reached by some examinees. This leads to a reduction in the size of the sample on which item analysis data for such items can be based; this reduction leads to an increase in the variance errors of difficulty and discrimination indices for these items. Thus, in some instances the reliability of the discrimination index for an item depends on the position of the item in the try-out test. The test constructor must keep this point in mind because if the number in the sample that has not read successive items becomes sufficiently large, the item analysis data that can be computed may become so unreliable as to be not worth obtaining.

Even more serious than this progressive reduction in the reliability of item analysis data for successive items in a highly speeded tryout test is the systematic bias that may characterize the difficulty and discrimination indices, particularly the latter. Data biased in this way can be gravely misleading except to the most sophisticated test constructors, and even for the latter they are inconvenient to use. This bias may be caused by a tendency which the writer has observed for some examinees to rush through a test, marking items almost at random when they come to the more diffi-

cult items. In general, these examinees tend to be those of low rather than of high ability. The result is that the group of examinees who actually attempt items near the end of a highly speeded test consists of the exceedingly able, well-informed examinees who work rapidly and accurately and the dull, poorly informed examinees who rush through the items, marking answers almost at random. Thus, the distribution of criterion scores for the group that attempts items near the end of a highly speeded test tends to become platykurtic. For this reason, item-criterion correlation coefficients tend to become inflated for these items. Both Wesman and Mollenkopf have published data showing that this does occur (*66; 61*). Mollenkopf writes,

> The evidence of this study indicates that the best persons (in terms of criterion scores) attempt the most items, and usually are successful. The even spread of scores down into the low range indicates that some persons of low and average ability also mark late items. Two possible explanations are suggested why these persons of low ability thus hurry through the test: (*a*) they do not recognize their own limitations, believing themselves to be more capable than they really are, and (*b*) they are test-wise individuals who hope to better their scores by attempting a great many items.

This interpretation agrees with the writer's and suggests another reason for telling examinees not to guess wildly and for using an appropriate correction for chance in the scoring of tryout tests.

It has sometimes been proposed, when product-moment or biserial correlation coefficients have been computed between item and criterion scores, that a correction be made for any alteration in range of criterion scores for examinees who actually reach items near the end of a speeded tryout test. If a test constructor decides to apply such a correction, he should be careful to make use of an appropriate procedure. The basic formula derived by Pearson (and discussed by Kelley [*101;* p. 430]) for correcting a product-moment correlation coefficient for restriction of range on the basis of one of the correlated variables is not applicable to biserial correlation coefficients of the type most commonly used for item analysis purposes. A procedure developed in 1946 by Gillman and Goode (*48*) may be found reasonably applicable.

If the writer's recommendations for estimating the number of examinees who fail to reach each successive item and for using this information in computing difficulty and discrimination indices are followed, a small progressive reduction in the number of cases on which the item-analysis data are based will not seriously bias the indices. If the total number in the sample is used or if a distinction is not made between examinees who

have read an item and refrained from marking it and those who have failed to reach the item in the time limit, the indices can be strongly biased merely because of the position of the item in the tryout form.

An example is provided by a five-choice item that was not reached in the time limit by approximately one-quarter of the sample. The data for the correct answer were:

	Number in Highest 27 Percent	Number in Lowest 27 Percent
Right	95	19
Wrong	2	6
Omit	2	28
Not reached	1	47
Total	100	100

Item analysis data obtained on the basis of three different formulas follow:

	Computing Formulas		
	$\frac{R - \frac{W}{4}}{N - NR}$	$\frac{R}{R + W}$	$\frac{R}{N}$
Item-test correlation coefficient read from Flanagan table	.69	.50	.76
Difficulty index (average of percents in high and low groups)	64	87	57

Criterion-Score Reliability

Any lack of reliability in the criterion variable affects all item-criterion coefficients in the same manner (though not by the same amount). Correction of these coefficients for lack of perfect reliability in the criterion will have no effect on their rank order, but it will render them all somewhat more unreliable. It will also impair estimates of the accuracy of prediction achieved by using them in combination. If, for some reason, the reliability of the criterion used for evaluating items in the tryout form of the test were altered considerably (without any change in the mental or physical skills measured) when the items were used for practical purposes, estimates of their prediction accuracy under the changed conditions could be secured by means of the following formula:

$$R_{ic} = r_{ic}\sqrt{\frac{R_{cc}}{r_{cc}}}, \tag{22}$$

where R_{ic} = the item-criterion correlation coefficient when the criterion reliability coefficient is R_{cc},

r_{cc} = the original criterion reliability coefficient,

R_{cc} = the altered criterion reliability coefficient,

r_{ic} = the item-criterion correlation coefficient when the criterion reliability coefficient is r_{cc}.

The Use of Item Analysis Data

Revision of Test Items

The previous discussion of item analysis data in this chapter has concerned the tabulation of right answers, wrong answers, and the number of examinees who did not reach each item. These tabulations are indeed sufficient to provide data for computing difficulty and discrimination indices, but they must be supplemented with choice-by-choice tabulations if the full value of item analysis techniques is to be realized. The exact form of these choice-by-choice tabulations depends on the type of statistics being used for difficulty and discrimination indices. Following is a choice-by-choice tabulation of the kind of data recommended by the writer. The item used for illustrative purposes is based on a reading passage that is not reproduced here.

	HIGH 27% Percent Selecting Choice (N = 100)	LOW 27% Percent Selecting Choice (N = 100)
13. Which one of the following words does not seem to go with the tone of the passage?		
A Monopoly (line 2)	6	13
B Renaissance (line 10)	16	20
C Lucky (line 12)	23	21
D Dozen (line 14)	6	10
*E Jackpot (line 22)	49	26
Omitted item	0	10
Did not reach item	0	0

* Correct answer.

It will be noted that 49 percent of the high group selected the correct answer to this item, choice E, while only 26 percent of the low group selected it. After these two percents have been corrected for chance success, they are used to enter the Flanagan table and to obtain the estimated percent of examinees in the entire sample who *know* the answer to the item. For purposes of obtaining information that will permit revising the item, we are as much interested in the percents in the high and low groups that select the incorrect choices as we are in the percents that select the correct choice. Note that incorrect choices A, B, and D are more attractive to the examinees in the low group than in the high group. This is just what the item writer hoped would be the case. Incorrect choice C, however, does not work so

well. A slightly larger percent of the high group than of the low group select it. Consideration of the reasons for this indicates the probability that examinees who are high in general reading ability regard "lucky" as a colloquial word. It isn't, of course—certainly not in the same way that "jackpot" is—but "lucky" does not serve as a discriminating incorrect choice, so the item would be improved by replacing "lucky" with a word that would constitute a more discriminating incorrect choice. Psychological insight and ingenuity would have to be exercised to secure a replacement. Careful consideration of the item analysis data and its relation to the passage on which the item is based would be required for this purpose. Comments of expert critics should be consulted to make sure that some previously undetected ambiguity in the item is not misleading the examinees in connection with choice C.

Inspection of choice-by-choice data of the kind illustrated above will ordinarily reveal many incorrect choices that are discriminative in the wrong direction and many that attract virtually no examinees. The former operate to destroy an item's discriminating power while the latter are nonfunctioning and may waste space and reading time. When a test constructor is provided with data of these kinds, he may delete choices that are grossly invalid and nonfunctioning. The efficiency of measurement of items that have been pruned in this way may be considerably improved.

Ordinarily, however, the test constructor will not be content to make so limited a use of the revealing data provided by a choice-by-choice item analysis. Guided by the hints offered in the data concerning the mental processes of the examinees, he can often replace invalid and nonfunctioning choices with a good chance of improving the efficiency of measurement of the item. Whether he has succeeded can be determined only by administering the revised items to samples like those used for the initial tryout. The expense and labor involved in tryouts or item analyses may make more than one tryout impracticable.

Occasionally, an incorrect choice that is markedly discriminative in the wrong direction represents a concept that cannot be removed without destroying the whole point of the item. This situation can arise as a result of testing a point about which considerable misinformation has been circulated. In the writer's opinion, if revision of an item involves destroying its point, the item either should be discarded or should be used without revision in spite of its probable lack of efficiency. An item should never be subjected to revision that destroys its point. Items that are not discriminative in the entire tryout sample will often be found capable of discriminating between the very best of the examinees and all of the others or between

the least capable and all of the others. Such items can be highly useful for specialized purposes.

A second illustration of the improvement in test items that may be obtained by editorial revision based on item analysis data is provided by the data in Table 7. Here we have detailed information showing the validity of each choice in an item before and after revision. In this case, the criterion was success or failure in learning to fly an airplane well enough to graduate from elementary flying training in the Army Air Forces. Note

TABLE 7

ITEM ANALYSIS DATA FOR A TEST ITEM BEFORE AND AFTER REVISION

Original Item*	Percent of Graduates Selecting Choice (N=259)	Percent of Eliminees Selecting Choice (N=84)	Revised Item*	Percent of Graduates Selecting Choice (N=200)	Percent of Eliminees Selecting Choice (N=200)
The valve cup is probably made of			The valve cup is probably made of		
A spring steel.	37	40	A celluloid.	1	2
*B composition rubber	49	49	*B composition rubber.	65	57
C cork.	1	1	C cork.	4	4
D tin.	1	2	D tin.	5	8
E bakelite.	7	7	E bakelite.	12	15
Did not reach	0	0	Did not reach	4	7
Omitted	5	1	Omitted	9	7
Percent answering correctly = 49			Percent answering correctly = 61		
r_{tet} with criterion = .00			r_{tet} with criterion = .10		

* Both original and revised items were based on the same diagram, which is not reproduced here. The asterisk in front of choice B indicates that it was the correct answer to each item.

that, in the original item, choices A and B attracted virtually all of the examinees and that the percentages of graduates and eliminees selecting choice B (the correct answer) were the same. Hence, the item had no correlation with the criterion. In an effort to improve the item, choice A was changed from "spring steel" to "celluloid." The latter constituted a far less attractive choice to similar examinees in the second tryout group. The distribution of answers changed in such a way that the item became slightly easier and displayed some validity. It cannot be said that the change was strikingly successful, but it was accompanied by an increase in the item's validity coefficient.

Needless to say, these data should not be considered as proof of anything. Too few cases are involved for that. But they do illustrate one important use of item analysis data. Test constructors accustomed to dealing with internal-consistency item analysis data may find the item-criterion

correlation of .10 very low compared with the values of .40 to .60 that they commonly obtain. It must be remembered, however, that such high values are rarely encountered when the criterion is not the total test score, and that they are almost never encountered when the criterion is a realistic one—such as performance in a course of training or a job.

It is interesting that many invalid distracters are found in items that have been carefully edited and checked by subject-matter experts. This emphasizes the well-known fact that because a distracter is discriminative in the wrong direction we cannot conclude that it is too nearly correct from a factual point of view. Conversely, the fact that an incorrect choice is too nearly a correct answer does not necessarily mean that it will turn out to be discriminative in the wrong direction when it is subjected to item analysis. Item analysis techniques cannot alone be relied upon to detect errors and ambiguities; expert criticism and editing are indispensable in test construction. The full value of item analysis techniques cannot be realized unless criticisms of the items by recognized authorities are available for reference.

The Relationship of Item Difficulty to Item Discriminating Power

The relationships of item difficulty and item discriminating power are rather complicated and little understood, but, since a thorough comprehension of their relationships is basic to an intelligent selection of test items, it is important that an explanation of the matter be presented here. The magnitude of a product-moment item-criterion correlation coefficient is dependent on two separate elements: the underlying relationship of the variables measured by the item and the criterion, and the number of discriminations the item can make among the members of a given sample. If we assume rectilinearity of regression, the underlying relationships can properly be estimated by statistics such as the biserial or tetrachoric correlations, or by short-cut procedures that lead to the use of the Flanagan table or the Davis item analysis chart. The number of possible discriminations can be determined from the percent of examinees who answer the item correctly.

When a product-moment correlation coefficient is computed between item and criterion scores, its value represents a combination of the influences produced by the two factors mentioned. As stated previously, its value rises as the underlying relationship of the variables increases and as the number of discriminations the item can make approaches its maximum at the difficulty level where 50 percent of the examinees know the answer. For a

given degree of underlying relationship between the item and criterion, the product-moment correlation is maximized at the 50 percent difficulty level. If the item is to be used alone or in combination with others for prediction purposes among examinees having the same average and distribution of ability as those in the tryout group, no other statistic (based on rectilinear regression) will serve so well as the product-moment coefficient. But if we wish to generalize the findings obtained in one group to another having a different average level of ability or a different point of dichotomy with respect to the criterion variable, we can improve on the results of using a product-moment coefficient in the original group by using data pertaining separately to the two factors entering into the product-moment *r;* namely, the underlying relationship and the difficulty level. If a single test item is answered correctly by 50 percent of a group of 100 examinees, it discriminates between the 50 who pass it and the 50 who fail it. The total number of discriminations made by the item is, therefore, 50 $\times$ 50, or 2,500. It is obvious that this is the largest number of discriminations that the item can make. Note that if 40 percent of the 100 examinees pass the item, it will make 40 $\times$ 60, or 2,400, discriminations. If 1 percent of the 100 examinees pass the item, it will make only 1 $\times$ 99, or 99, discriminations. These data illustrate the fact that if we were to build a test consisting of one item (which we never do), it should be of 50 percent difficulty for the sample in which it is to be used if it is to be maximally discriminative. But notice that a single-item test is useless if we want to discriminate between the ability of two examinees who fail it. If we want to discriminate between examinees capable of passing an item at the 30 percent difficulty level and those not capable of doing so, we have to employ an item of 30 percent difficulty level. This item will make only 30 $\times$ 70, or 2,100, discriminations in the sample, but they will be useful discriminations. These data indicate that the purpose for which a test is to be used is more important in determining the level of difficulty of its component items than are other considerations.

Richardson has shown that if one wants to construct a test to differentiate examinees above a given level of ability from those below that level of ability (without making any distinctions among examinees in the two groups), all of the items used in the test should be of a difficulty level such that they will be marked correctly by half of the examinees *at the level of ability represented by the line of demarcation* (82). His mathematical demonstration of the point fits in nicely with the practical illustration provided above. This is an important consideration in selecting test items because examinations are sometimes built for the purpose of separating

examinees into two groups. Selection tests used by industrial organizations are often used with passing marks. For a test of this kind we should select items at a certain difficulty level so that the maximum discriminating power of the test will be exerted at the passing mark. The items will all be of 50 percent difficulty only when we wish to exclude 50 percent of the examinees. Items of 50 percent difficulty level have, in general, no peculiar merit.

Most examinations used in schools are not employed mainly to divide students into two groups at the passing mark. Ordinarily, distinctions must be made among passers and failers in order to assign marks (say, A, B, C, D, and E). Discriminations must be made throughout the range of scores to rank students in order of ability, though better-than-average discrimination power should, theoretically, be exerted at the dividing points between the scores assigned A and B, B and C, etc.

Suppose that we now consider a test of 10 items to be used with a group of 100 examinees. If the items are all uncorrelated with one another, the number of discriminations will be maximized when the items are all of 50 percent difficulty. If all items were perfectly correlated (and thus perfectly reliable), the number of discriminations made by *10* items at 50 percent difficulty level would be identical with the number of discriminations made by *one* item of 50 percent difficulty. The maximum number of discriminations that could be made if the 10 items were all at one level of difficulty and perfectly intercorrelated would be 2,500, but if we spread the 10 items at intervals of 9.09 percent from 9.09 percent to 90.90 percent, 4,562 discriminations could be made. The latter arrangement would be optimal for 10 items under the circumstances specified.

These data suggest that maximum discrimination among all the members of a group may be obtained when the items in a test are uncorrelated by using items all of 50 percent difficulty, but that when the items are perfectly correlated, the items should be spread over the range of difficulty. However, the items in a test are never found to be either wholly uncorrelated or perfectly correlated. Hence, the limiting cases we have used for illustrations serve merely to guide our thinking about the distribution of item difficulties required to obtain maximum discrimination throughout the entire range of scores in a given sample. The analytic solution to this problem for any given level of item difficulty and item intercorrelation has been discussed by the writer, but the computational labor required to provide a practical guide for the test constructor has not yet been performed (*72*). We can see intuitively, however, that since many kinds of test items have low intercorrelations, a distribution of item difficulties

clustered around the 50 percent level would often approximate the distribution required to obtain maximum discrimination throughout the range of scores. For vocabulary items and other types that tend to have relatively high intercorrelations, the distribution of difficulty indices should be made more platykurtic than usual if equal accuracy of measurement and maximum discrimination are desired throughout the range of scores.

It should be made clear at this point that there is often good reason for avoiding a distribution of item difficulties that will cause a test to yield equal accuracy of measurement throughout the range. The separation of a group of examinees into two subgroups calls for all possible accuracy of measurement at the dividing line; the assignment of marks (which calls for the division of a group into several parts) demands maximum accuracy of measurement at the several dividing points scattered along the range of scores. Even when the members of a group are to be placed in rank order, the purpose for which the rank order is intended usually dictates a portion of the range that deserves greater accuracy of measurement than another. For example, in selecting teachers for a large school system, the qualifying examination should yield greater accuracy in the rank order of the applicants who obtain high scores since it is unlikely that applicants who obtain low scores will be given serious consideration unless there is a grave shortage of applicants for teaching positions. To select a half-dozen high school graduates to be given university scholarships calls for extraordinary accuracy of the rank order at the extreme upper end of the distribution of scores. A test designed for this purpose should be made up almost entirely of exceptionally difficult items in each of the fields measured.[15]

Brogden has presented unusually interesting data regarding the effect on test validity of deliberate control of item difficulty and intercorrelation (*69*). These confirm the somewhat theoretical formulation presented above. For example, from Brogden's data we find that if a test of 45 items (all of 50 percent difficulty and having tetrachoric intercorrelations of .60) has a correlation with a criterion variable of .950, we can increase this correlation to .961 by adding 108 similar items. But we can obtain a correlation of .962 by using only 45 items like the originals except that their difficulty indices form a rectangular instead of a point distribution. Thus, under these unusual conditions, a test of 45 items fairly well adjusted for difficulty level may be as useful as a test of 153 items all of 50 percent difficulty.

From the preceding discussion it is apparent that the distribution of difficulty indices should be controlled very closely if maximum efficiency

[15] For a rigorous treatment of this issue see "Selected References," p. 327, item *85*.

in measurement is to be attained. When this direct control is applied by means of difficulty indices, it is desirable to select those items of suitable difficulty that have high discrimination indices and that meet the approval of subject-matter specialists. For this purpose, discrimination indices that reflect the underlying relationships of the item and criterion variables are to be preferred to product-moment coefficients.

Principles of Selecting Items for Specific Types of Tests

For purposes of this discussion all tests may be divided into two groups which we shall label "self-defining tests" and "predictor tests." The self-defining test is so named because the weighted sum of the skills and abilities measured by it actually defines what it measures. Achievement tests or conventional intelligence tests are common examples of self-defining tests. More unusual self-defining tests are Professor Thurstone's Tests of Primary Mental Abilities. All of these tests are constructed to measure a combination of skills and abilities specified in advance by the test constructor. A great deal of care and skill is exercised to make sure that the items included in a self-defining test actually measure what the test constructor wants the test to measure, but in the last analysis it is self-defining. It should be noted that self-defining tests are often used for prediction purposes, in which case they may be regarded as predictor tests.

A predictor test is one that is constructed on the basis of empirical data to correlate as highly as possible with a criterion variable.[16] Once the items for a predictor test have been tried out, selection of the items for the final form may, theoretically, be determined by multiple regression techniques or an approximation of them. The subjective judgment of the test constructor in the selection of items is reduced to a minimum.

Self-defining tests

The power test.—A power test is intended to measure level of performance with respect to some defined skill or ability, or some specified combination of skills and abilities. The first step in constructing any self-defining test is to define the content to be measured in such a way that recognized authorities in the field will agree that the definition is adequate. This definition is best prepared in the form of an outline in which each topic is weighted in proportion to its importance, as judged by analyses of textbooks and curriculums and the pooled opinions of experts (see chapter 6).

[16] A suppressor test is a special kind of predictor test that is constructed to have a high correlation with an existing predictor and a low correlation with the criterion; see "Selected References," p. 328, item *104*.

If the items in a tryout test have been well constructed and are well apportioned among the skills and abilities to be measured, the total score may be regarded as a valid, if perhaps inefficient, measure of the intended subject-matter field. In other words, the test will be regarded as a valid measure by experts qualified to make such judgments. Given a tryout test of this kind, selection of items for the final form may be accomplished as follows:

1. A difficulty index should be computed for each item.

2. A discrimination index should be computed for each item, preferably an index that reflects the underlying item-criterion relationship. Since the total test score is used as the criterion, biserial *r* may be computed or one of the tables mentioned previously may be employed.

3. Consideration should be given to possible differences in the average level of ability between the tryout group and various groups in which the final test is likely to be used. If difficulty indices are read from the Davis table, any difference that is presumed to exist may be expressed as a constant number of points of difficulty.

4. The number of items desired at each level of difficulty should be estimated. The shape of this distribution will depend on the purpose of the test and on the degree of homogeneity of the items. If the test is to be used with a "passing mark" and no further distinctions are to be made, all items should be of an appropriate difficulty level. If 30 percent of the applicants of ability corresponding to the predetermined average level of difficulty are to be rejected, the items should all be of a difficulty level corresponding to that represented by a difficulty index of 70 percent among the applicants. This procedure assumes that the underlying item-criterion relationships are equal. In practice, items above and below this level will have to be used since not enough items of precisely the desired level of difficulty are likely to be obtained that meet other requirements for selection.

To obtain maximum discrimination among all the examinees, the shape of the distribution of items around the 50 percent level (among the group in which the test is to be used) should be made leptokurtic if the items are rather heterogeneous in content and platykurtic if the items are homogeneous. This procedure will tend to maximize the correlation between obtained and true scores on the final form of the test; for self-defining tests, this correlation may be regarded as the validity coefficient of the test. The degree of homogeneity of well-edited items can be estimated by the average magnitude of the discrimination indices. An average biserial *r* of .60 to .70 (when the number of items in the criterion score is about 100) may be considered indicative of rather homogeneous items. The correspond-

ing limits for Davis discrimination indices are about 40 to 55. An average biserial r of .20 to .30 (when the criterion score includes about 100 items) is characteristic of rather heterogeneous items. The corresponding limits for Davis discrimination indices are 12 to 19.

If considerably greater accuracy of measurement is desired in one part of the range of difficulty than in another, items should be concentrated in that part of range.

5. All the tryout items should next be separated into groups indicated in the outline of the test. From each separate group, a number of items should be selected tentatively that will be roughly proportional to the weight given each division in the test outline and that will create approximately the proper distribution of item difficulty indices. At the same time, an effort should be made to choose those items having the highest discrimination indices. Needless to say, any item that has not been approved by subject-matter experts should be excluded from consideration.

This is a complicated process; compromises have to be made within each division of the test outline, though sometimes compensating compromises can be arranged in other divisions. After long experience, the test constructor gets a "feel" for the process and learns how much compromise is either necessary or permissible under the particular circumstances. The test constructor cannot be a perfectionist when he is selecting items.

6. The entire group of test items should be read over as a unit to detect unnoticed overlappings of choices and to prevent cross-keying of items; that is, having the statement or choices in one item give a clue to the answer to another item.

7. The choice-by-choice item analysis data for each item should be studied and changes made in choices that are likely, in the test constructor's judgment, to improve the item's efficiency.[17] Ordinarily, revision of items in this manner may be expected to lower the difficulty index (or increase the difficulty) of an item in which an unattractive incorrect choice is replaced with one judged to be more attractive to examinees who obtain low scores on the criterion variable and to raise the difficulty index (or decrease the difficulty) of an item in which nondiscriminating choices are replaced with ones judged to be less attractive to examinees who obtain high scores on the criterion variable. The test constructor should try to make allowances for the estimated net effect of these adjustments in item difficulty.

8. The items should be grouped appropriately, arranged in order of

[17] These changes should ordinarily be made even though a subsequent tryout will not be conducted to ascertain their efficiency. It would be wasteful to refrain from making the maximum use of all available data in revising items even though a certain risk may be involved. Whenever possible, of course, a second tryout of revised items should be made.

difficulty, and have choices transposed to provide roughly an equal number of correct responses for each choice number or letter.

It may seem surprising that no more emphasis is placed on the use of item discrimination indices than is indicated by the procedures described above. The writer believes, however, that their importance has often been greatly overemphasized in the construction of self-defining tests. Broadly speaking, they should be given more weight when the items in the tryout test prove to be extremely homogeneous. As the items making up the criterion score become less homogeneous, less weight should be given to the discrimination indices in selecting items.

The speed test.—A speed test should, strictly, comprise items that are all so easy that, given time, all examinees could answer them correctly. Thus, scores derived from a speed test will reflect mainly the rapidity with which examinees answer the items. The selection of items for speed tests is sometimes based partly on data provided by administering the items with sufficient time so that every examinee has a chance to try every item. Items of the desired level of difficulty are identified in this manner.

The mastery test.—A mastery test is a special kind of test used when it is desired to determine the proportion of essential subject matter a pupil has learned. The content is determined by the judgment of experts, as in the case of ordinary tests. However, since the purpose is to measure only abilities or skills that should have been mastered by every pupil, the items are necessarily extremely easy and difficulty indices are not used to exclude items that are answered correctly by 100 percent, or almost 100 percent, of the tryout groups. Such items do not discriminate among the individuals tested and are ordinarily omitted from the final form of a test used to rank individuals rather than to ascertain their mastery of certain subject matter. Correlations between success or failure on each item and the criterion score are necessarily zero if everyone answers each correctly. Furthermore, these correlations are not very meaningful for items that are answered correctly by almost everyone. Hence, item analysis data are not especially serviceable for selecting items for mastery tests.

Predictor tests

The simple prediction test.—A simple prediction test is one used to provide an estimate of the rank order of a group of examinees with respect to a designated criterion variable. For example, a civil service examination in shorthand is intended to rank the applicants for positions as stenographers with respect to their expected facility in taking dictation on the job.

To select the best combination of items for a simple selection test, we

need data showing the correlation of each type of item with the criterion and with the other types of items that have been tried out. The types of items that have higher-than-average correlations with the criterion and lower-than-average correlations with one another represent the most promising sources of items for the final form of the test. Consequently, we select from the proper groups those individual items that are of the desired level of difficulty, that are judged satisfactory by subject-matter experts, and that have the highest correlations with the criterion. If item analysis data are available with the total score on all the items, some preference should be given to items that have the lowest internal-consistency discrimination indices. To check on the validity of the resulting final form, an entirely new sample of examinees must be tested. The importance of checking validity coefficients on a sample different from that used in constructing the test cannot be overemphasized.

In actual practice, the selection of items for a simple prediction test may be accomplished efficiently by obtaining two sets of item discrimination indices—one with the total score of the test in which the item has been tried employed as the criterion variable and one with the variable to be predicted employed as the criterion. If the test is to be used with groups in which the distribution and average level of ability are essentially the same as in the tryout samples, the item discrimination indices may be expressed as product-moment correlation coefficients and items should be chosen to minimize the average of the internal-consistency coefficients and maximize the average of the coefficients with the variable to be predicted. The procedures described by Horst (*93*), Flanagan (*88*), or by Richardson and Adkins (*96*) are reasonably effective methods of accomplishing this. Gulliksen has reported data concerning the validity of a test constructed in accordance with Horst's technique (*89*).

If what is desired is minimum over-all error of prediction in the least-squares sense, the difficulty indices of items chosen in accordance with Horst's technique need not be considered at all under the circumstances specified. Often, however, the group with which the test is to be used is not of the same average level of ability as the group in which tryout data were obtained. Furthermore, the test constructor is not always interested in minimizing over-all errors in prediction; he may wish to predict scores within a certain part of the range, or at a particular point, with great accuracy even if this means he must accept rather unreliable predictions in other ranges of scores. To accomplish any of these objectives, the distribution of item difficulty indices must be deliberately controlled; items must be selected for the final form in such a way as to obtain the desired distribu-

tion of item difficulty indices, minimize item-test relationships, and maximize item-criterion relationships. These relationships should be expressed as biserial r's, tetrachoric r's, Flanagan r's, or Davis discrimination indices.

Whenever item analysis data are used to select items with minimum internal consistency, great care must be exercised to make sure that the low internal-consistency discrimination indices found for some items are the result of low relationships between the underlying variables measured by the items and by the total score used as a criterion and not the result of some ambiguity or other fault in the items themselves. To make sure of this, the criticisms of subject-matter experts should be consulted, and careful scrutiny of the items should be made.

The effectiveness of selecting items by means of the procedure recommended for simple prediction tests depends on the following factors: first, the reliability of the discrimination indices; and second, the degree of correlation between the total test scores used as a criterion for the internal-consistency analysis and the variable to be predicted. If the indices are not highly reliable, selection of items having low internal-consistency discrimination indices and high validity indices becomes largely a matter of chance and cannot be justified. The use of large samples of examinees is imperative when indices are to be used for this purpose. Samples of 1,000 examinees are unquestionably minimal.

As the correlation between the total test score used as the criterion for the internal-consistency analysis and the variable to be predicted increases, it becomes less important to use internal-consistency data as a supplement to item-criterion data. The lower this correlation, the greater the gain in test validity that becomes possible by using two sets rather than one set of item analysis data.

Some psychologists have expressed concern that the procedure for maximizing test validity recommended in this section will lead to an undesirable degree of heterogeneity in tests. There is no doubt that the procedure tends to increase item heterogeneity, but it also tends to maximize test validity. Since the latter is usually the proper standard by which the merit of a simple prediction test is to be judged, it seems hardly justifiable to condemn the heterogeneity of the items that leads to high validity. In actual practice, items must be selected from the tryout form of the test, which includes items judged to be appropriate on the basis of the test outline. Therefore, the advantageous heterogeneity produced by selecting items with low intercorrelations is held within close limits.

The simple selection test.—A test used for simple selection is only a special type of simple prediction test. Instead of ranking individuals in

terms of expected performance with respect to a criterion variable, a simple selection test is designed to separate examinees into two groups—those to be accepted and those to be rejected. No further distinctions are to be made within the two groups. The methods used for selecting items are those appropriate in the case of simple prediction tests except that items are chosen so that their difficulty indices will be as close as possible to the level of difficulty represented by the line of demarcation between the two groups to be formed. In other words, every item should be as nearly as possible of 50 percent difficulty *for examinees at the level of ability represented by the "passing mark."*

The multiple-prediction test.—A multiple-prediction test is used to obtain separate rank orders of performance for a group of examinees with respect to more than one criterion variable. Unless the various criteria to be predicted are highly intercorrelated, the multiple-prediction test must be composed of separately scored subtests. These subtests are correlated with each other and with each one of the variables to be predicted, and the effective weights capable of minimizing the over-all error of prediction for each criterion variable are assigned to the subtests.

For each subtest, items should be selected that measure the same mental function. Each subtest should approximate a "pure" test, so internal-consistency item analysis data should be used to select items for each subtest that have high correlations with the total score on that subtest and, if possible, low correlations with the total scores on other subtests. A description of a test constructed in this manner was presented by Davis in 1947 (*43,* pp. 92-95). Theoretically, there is no reason why "pure" tests must have zero or low intercorrelations, as L. L. Thurstone has so well pointed out on many occasions. Nonetheless, there is some practical efficiency achieved by using item analysis data to minimize the intercorrelations of subtests employed to obtain weighted composite scores.

The shape of the distribution of difficulty indices that is optimal for each subtest depends on the degree of intercorrelation of its component items, on the degree of intercorrelation of the subtest scores included in the composite, on the statistic used to express the indices, and on the purpose for which the composite scores are to be used. A detailed exposition of this matter is too long to be included in this chapter, but it can be said that, in general, the accuracy of measurement of scores in each subtest should be equal over a rather wide range of ability because subtests are generally selected partly on the basis of low intercorrelations; hence, a given individual's weighted composite score is apt to depend on subtest scores representing rather different levels of ability.

To obtain subtest scores having equal accuracy of measurement over a rather wide range, the principles recommended for this purpose in the preceding section on simple prediction tests should be followed. Since the items in a subtest are selected largely on the basis of homogeneity of content, their difficulty indices should cover a wide range rather than cluster close to the 50 percent level of difficulty. The use of only items of 50 percent difficulty is more inappropriate than usual in the case of a so-called "pure" test unless the test is used with a "passing mark" at the level of ability represented by a difficulty index of 50.

The multiple-selection test.—Instead of providing separate rank orders for several criterion variables, the multiple-selection test is designed to separate the examinees into only two groups with respect to each criterion. It is analogous to the simple selection test; therefore, accuracy of measurement of the weighted composite scores derived from its subtests should be maximized at the several lines of demarcation. To accomplish this as effectively as in the case of a simple selection test is impossible. Reduction in the intercorrelations of the subtests lowers the extent to which accuracy of measurement in the composite scores can be concentrated at the line of demarcation. In practice, one can often do little better than utilize the principles of selecting items recommended for use with multiple-prediction tests.

The differential classification test.—A differential classification test is used to determine (at stated levels of confidence) in which one of several criterion variables any given person will show the highest level of competence. For example, if a man *must* be assigned to one of four available jobs, we use a differential classification test to determine in which one of the jobs he is likely to be most successful. Our objective in differential classification, then, is not to determine how well he is likely to be able to perform any skill that the four jobs have in common, but rather to determine how well he is likely to perform the skills that are unique to each job. This means that an item should be assigned to one of the subtests in the differential classification test on the basis of differences between its correlations with the several criteria to be predicted and without regard for internal-consistency data. The larger the difference between the correlation of an item with one criterion and the median of its correlations with other criteria, the better the item. (The comparison of differences between correlation coefficients should be made only after their transformation into Fisher's z or Davis' discrimination index.) An excellent item for purposes of differential classification would be one that displayed a correlation of .40 with one job criterion and correlations of .11, .02, and —.07 with the

other job criteria. There is no mathematical necessity that the subtests constructed in this fashion be "pure" tests, but the method used will tend to lower their intercorrelations.

It is interesting to note that the common variance of the criterion variables is irrelevant for differential classification; hence, if factorial studies are performed on matrices made up entirely of them, it is best to analyze either the total variance or the non-chance variance. If communalities are used, the variance to be predicted for purposes of differential classification is likely to have been excluded in advance from the analysis.

In practice, genuine differential classification tests are rarely sought because the scores resulting from their use should predict to a large degree only unique elements of each criterion variable. If a man were assigned to a certain job because his subtest scores on a differential classification test showed that he was more likely to succeed in it than in any other of the criterion jobs, he might still not perform satisfactorily because his absolute level of ability in the combined common and unique variance of the job might be too low. The differential classification test scores would have indicated only his relative standing in the criterion jobs; they would have indicated very little about his actual level of performance.

This means that differential classification tests are appropriate when every examinee must be assigned to work in one of the criterion jobs or when all of the examinees have previously been selected by means of some sort of qualifying examination, perhaps one that measures largely the common variance of the criterion jobs. If some testees are to be rejected for all of the jobs or if certain absolute levels of performance are set as minimal for acceptance in one or more of the criterion jobs, a multiple-selection test is used for the purpose. A differential classification test is likely to be of most social utility in time of national mobilization. Even in that situation, however, not everyone called up for national service is likely to be accepted.

Recording Item Data

The selection of items for the final form of a test can be greatly facilitated if the data that must be used are available in convenient form. Because the selection process usually involves combining items in several different ways before a satisfactory arrangement is found, it is most convenient to have each item, together with all data pertaining to it, on a single filing card. Specially printed 5×8 cards are quite suitable for this purpose; they are large enough to carry all the data that are ordinarily required and small enough to be manipulated easily. Figure 5 shows one

Test: Booklet 27 Part: II			Test: Booklet 27 Part: II		
Item no. 50 Author:			Item no. 50 Editor:		
Sample: Unclassified A/C			Sample: Classified as pilot		
Date:			Date:		
	High	Low		Graduate	Eliminee
N	100	100		582	175
A	87	65		71	65
B	8	14		12	18
C	3	11		9	9
D	1	8		6	5
E	0	0		1	1
Omit	1	2		2	3
NR	0	0		0	0
Corr.%	84	57			
Diff. Index	70			63	
r	.32			.12	

Item

50. A feathered propeller is used to

50-A cut down propeller drag.
50-B increase propeller thrust at high altitudes.
50-C secure idling motor speeds prior to landing.
50-D start an engine which has failed in flight.
50-E prevent glare.

FIG. 5.—Item data card (front)

Test:__________ Part:_____
Item no._____ Editor:_____
Sample:______________
Date:________________

	High	Low
N		
A		
B		
C		
D		
E		
Omit		
NR		
Corr.%		
Diff. Index		
r		

Test:__________ Part:_____
Item no._____ Editor:_____
Sample:______________
Date:________________

	High	Low

Comments

Suggest changing item to read:

An aircraft propeller is feathered to

A cut down drag.
B increase its thrust at high altitudes.
C make the engine idle properly during landings.
D start the engine if it fails during flight
E prevent ice from forming on it.

FIG. 5 (*Cont.*).—Item data card (reverse side)

side of such a card with an item and two sets of item analysis data pertaining to it already entered on the card. The item was cut out of the tryout booklet (which had been photo-offset) and was affixed with rubber cement to the card in the space provided. This procedure saved the labor of retyping and proofreading each item. If the number of lines in an item is so large that it will not fit in the space provided, the top part can be pasted and the bottom part folded over.

The choice-by-choice data derived from an internal-consistency item analysis are shown in the two columns at the extreme left of the card. Similar data derived from an item analysis, using an external criterion (graduation or elimination from elementary pilot training in the Army Air Forces), are shown in the two right-hand columns. The difficulty index in each of the two samples is shown as the corrected percent of examinees who marked the right answer. There was no one who failed to reach the item in the time limit, though a small percent of each group omitted the item after reading it. The discrimination index derived from the internal-consistency analysis is expressed as a correlation coefficient read from the Flanagan table; the one derived from the external-criterion analysis is expressed as a tetrachoric correlation coefficient read from computing diagrams.

The reverse side of the same item card is shown in Figure 5 *continued.* Space is provided for additional item analysis data and for comments. A revision of the item has been suggested by one of the critics to whom it was submitted for scrutiny. The wording has been improved and choice E has been replaced with one that appears more attractive to examinees, especially to those likely to fail in pilot training.

If sufficient clerical help is available, it is desirable to make two cards for each item that is tried out. The first copy is filed by subject matter; the second is used to construct the final form of the test and is filed numerically by test form if it is included in a final form. If not, it is filed by subject matter in a "discarded item" file that can be used as a source of ideas for new items. In any large-scale test construction agency an item file of the kind described will not be found more elaborate than is desirable. For teachers and individuals who construct occasional tests, a single file of item cards may be adequate.

Selected References

Annotated lists of references pertaining to item analysis data and techniques of item selection may be found in the compilations by Swineford and Holzinger published in the June issue of *The School Review* for several years past. Special issues of the *Review of Educational Research,* the first and second issues of the

Yearbook of Research and Statistical Methodology edited by Buros, and the annual bibliographies published by Good in the *Journal of Educational Research* may also be consulted for classified lists of references concerning test construction.

In view of the bibliographical material already available, no effort has been made to present a complete list of references here. Only articles of special interest in connection with this chapter are listed.

Techniques of Item Analysis

1. ADKINS, D. C., *et al. Construction and Analysis of Achievement Tests: The Development of Written and Performance Tests of Achievement for Predicting Job Performance of Public Personnel.* Washington: Government Printing Office, 1947.
2. ARNOLD, J. N. "Nomogram for Determining Validity of Test Items," *Journal of Educational Psychology,* **26**: 151–53, 1935.
3. BAKER, K. H. "Item Validity by the Analysis of Variance; An Outline of Method," *Psychological Record,* **3**: 242–48, 1939.
4. BARTHELMESS, H. M. *The Validity of Intelligence Test Elements.* ("Teachers College Contributions to Education," No. 505.) New York: Teachers College, Columbia University, 1931.
5. CHAPANIS, A. "Notes on the Rapid Calculation of Item Validities," *Journal of Educational Psychology,* **32**: 207–304, 1941.
6. CLARK, E. L. "A Method of Evaluating the Units of a Test," *Journal of Educational Psychology,* **19**: 263–65, 1928.
7. DAVIS, F. B. *Item-Analysis Data: Their Computation, Interpretation, and Use in Test Construction.* ("Harvard Education Papers," No. 2.) Cambridge: Graduate School of Education, Harvard University, 1946.
8. DUBOIS, P. H. "A Note on the Computation of Biserial *r* in Item Validation," *Psychometrika,* **7**: 143–46, 1942.
9. DUNLAP, J. W. "Nomograph for Computing Bi-Serial Correlations," *Psychometrika,* **1**: 59–60, 1936.
10. ———. "Note on Computation of Biserial Correlations in Item Evaluation," *Psychometrika,* **1**: 51–58, 1936.
11. FERGUSON, G. A. "Item Selection by the Constant Process," *Psychometrika,* **7**: 19–29, 1942.
12. FLANAGAN, J. C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distributions," *Journal of Educational Psychology,* **30**: 674–80, 1939.
13. FULCHER, J. S., and ZUBIN, J. "The Item Analyzer: A Mechanical Device for Treating the Four-fold Table in Large Samples," *Journal of Applied Psychology,* **26**: 511–22, 1942.
14. GARLOUGH, L. N. "A Convenient Method for Calculating Indices of Ease and of Differentiating Ability for Individual Test Questions," *Journal of Educational Research,* **35**: 611–17, 1942.
15. GOHEEN, H. W., and KAVRUCK, S. "A Worksheet for Tetrachoric *r* and Standard Error of Tetrachoric *r* Using Hayes' Diagrams and Tables," *Psychometrika,* **13**: 279–80, 1948.
16. GUILFORD, J. P. "The Phi Coefficient and Chi Square as Indices of Item Validity," *Psychometrika,* **6**: 11–19, 1941.
17. GUTTMAN, L "The Cornell Technique for Scale and Intensity Analysis," *Educational and Psychological Measurement,* **7**: 247–80, 1947.
18. JOHNSON, A. P. "An Index of Item Validity Providing a Correction for Chance Success," *Psychometrika,* **12**: 51–58, 1947.
19. KELLEY, T. L. "The Selection of Upper and Lower Groups for the Validation of Test Items," *Journal of Educational Psychology,* **30**: 17–24, 1939.
20. KOLBE, L. E., and EDGERTON, H. A. "A Table for Computing Biserial *r*," *Journal of Experimental Education,* **4**: 245–51, 1936.

21. KUDER, G. F. "Nomograph for Point Biserial *r*, Biserial *r*, and Four-fold Correlations," *Psychometrika*, 2: 135–38, 1937.
22. LAWSHE, C. H., JR. "A Nomograph for Estimating the Validity of Test Items," *Journal of Applied Psychology*, 26: 846–49, 1942.
23. LENTZ, T. F., and WHITMER, E. F. "Item Synonymization: A Method for Determining the Total Meaning of Pencil-Paper Reactions," *Psychometrika*, 6: 131–39, 1941.
24. LEV, J. "Evaluation of Test Items by the Method of Analysis of Variance," *Journal of Educational Psychology*, 29: 623–30, 1938.
25. LINDQUIST, E. F., and COOK, W. W. "Experimental Procedures in Test Evaluation," *Journal of Experimental Education*, 1: 163–85, 1933.
26. LOEVINGER, J. *A Systematic Approach to the Construction and Evaluation of Tests of Ability.* ("Psychological Monographs," No. 285.) Washington: American Psychological Association, 1947.
27. LONG, J. A., and SANDIFORD, P., *et al. The Validation of Test Items.* Toronto: Department of Educational Research, University of Toronto, 1935.
28. MOSIER, C. I., and MCQUITTY, J. V. "Methods of Item Validation and Abacs for Item-Test Correlation and Critical Ratio of Upper-Lower Difference," *Psychometrika*, 5: 57–65, 1940.
29. RICHARDSON, M. W. "Notes on the Rationale of Item Analysis," *Psychometrika*, 1: 69–76, 1936.
30. RICHARDSON, M. W., and STALNAKER, J. H. "A Note on the Use of Biserial *r* in Test Research," *Journal of General Psychology*, 8: 463–65, 1933.
31. ROYER, E. B. "A Machine Method for Computing the Biserial Correlation Coefficient in Item Validation" *Psychometrika*, 6: 55–59, 1941.
32. TURNBULL, W. W. "A Normalized Graphic Method of Item Analysis," *Journal of Educational Psychology*, 37: 129–41, 1946.
33. VERNON, P. E. "Indices of Item Consistency and Validity," *British Journal of Psychology*, Statistical Section, 1: 152–66, 1948.
34. VOTAW, D. F. "Graphical Determination of Probable Error in Validation of Test Items," *Journal of Educational Psychology*, 26: 682–86, 1933.
35. ———. "Notes on the Validation of Test Items by Comparison of Widely Spaced Groups," *Journal of Educational Psychology*, 25: 185–91, 1934.
36. WALKER, H. M. "Item Selection by Sequential Sampling," *Teachers College Record*, 50: 404–9, 1949.
37. WALKER, H. M., and COHEN, S. *Probability Tables for Item Analysis by Means of Sequential Sampling.* New York: Bureau of Publications, Teachers College, Columbia University, 1949.
38. ZUBIN, J. "The Method of Internal Consistency for Selecting Test Items," *Journal of Educational Psychology*, 25: 345–56, 1934.

STUDIES RELATED TO ITEM ANALYSIS TECHNIQUES

39. BARRY, R. F. "An Analysis of Some New Statistical Methods for Selecting Test Items," *Journal of Experimental Education*, 7: 221-28, 1939.
40. BROGDEN, H. E. "On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency," *Journal of Educational Psychology*, 37: 65–76, 1946.
41. CARTER, H. D. "How Reliable Are the Common Measures of Difficulty and Validity of Objective Test Items?" *Journal of Psychology*, 13: 31–39, 1942.
42. DAVIS, F. B. "Notes on Test Construction: The Reliability of Item-Analysis Data," *Journal of Educational Psychology*, 37: 385–90, 1946.
43. ——— (ed.). *The AAF Qualifying Examination.* ("Aviation Psychology Program Research Reports," No. 6.) Washington: Government Printing Office, 1947. Appendix A.
44. DOPPELT, J. E., and POTTS, E. M. "The Constancy of Item-Test Correlation Coefficients Computed from Upper and Lower Groups," *American Psychologist*, 3: 261, 1948. (Abstract.)
45. FLANAGAN, J. C. *Factor Analysis in the Study of Personality.* Stanford, Calif.: Stanford University Press, 1935.

46. Forlano, G., and Pintner, R. "Selection of Upper and Lower Groups for Item Validation," *Journal of Educational Psychology,* **32**: 544–49, 1941.

47. Gibbons, C. C. "The Predictive Value of the Most Valid Items of an Examination," *Journal of Educational Psychology,* **31**: 616–21, 1940.

48. Gillman, L., and Goode, H. H. "An Estimate of the Correlation Coefficient of a Bivariate Normal Population When X Is Truncated and Y Is Dichotomized," *Harvard Educational Review,* **16**: 52–55, 1946.

49. Guilford, J. P., and Lacey, J. I. (eds.). *Printed Classification Tests.* ("Aviation Psychology Program Research Reports," No. 5.) Washington: Government Printing Office, 1947.

50. Handy, U., and Lentz, T. F. "Item Value and Test Reliability," *Journal of Educational Psychology,* **25**: 703–8, 1934.

51. Hayes, S. P., Jr. "Diagrams for Computing Tetrachoric Correlation Coefficients from Percentage Differences," *Psychometrika,* **11**: 163–72, 1946.

52. Horst, A. P. "Regression Weights as a Function of Test Length," *Psychometrika,* **13**: 125–34, 1948.

53. Jurgensen, C. E. "Table for Determining Phi Coefficients," *Psychometrika,* **12**: 17–29, 1947.

54. Katzell, R. A., and Cureton, E. E. "Biserial Correlation and Prediction," *Journal of Psychology,* **24**: 273–78, 1947.

55. Kroll, A. "Item Validity as a Factor in Test Validity," *Journal of Educational Psychology,* **31**: 425–36, 1940.

56. Lawshe, C. H., Jr., and Thayer, J. S. "Studies in Item Analysis: I. The Effect of Two Methods of Item Validation on Test Reliability," *Journal of Applied Psychology,*
31. ~~**21**:~~ 271–77, 1947.

57. Lentz, T. F.; Hirshstein, B.; and Finch, F. H. "Evaluation of Methods of Evaluating Test Items," *Journal of Educational Psychology,* **23**: 344–50, 1932.

58. Lord, F. M. "Reliability of Multiple-Choice Tests as a Function of Number of Choices Per Item," *Journal of Educational Psychology,* **35**: 175–80, 1944.

59. McNamara, W. J., and Weitzman, E. "The Economy of Item Analysis with the IBM Graphic Item Counter," *Journal of Applied Psychology,* **30**: 84–90, 1946.

60. Merrill, W. W., Jr. "Sampling Theory in Item Analysis," *Psychometrika,* **2**: 215–23, 1937.

61. Mollenkopf, W. G. "An Experimental Study of the Effects on Item-Analysis Data of Changing Item Placement and Test Time Limit," *Psychometrika,* **15**: 291–315, 1950.

62. Mosier, C. I. "A Note on Item Analysis and the Criterion of Internal Consistency," *Psychometrika,* **1**: 275–82, 1936.

63. Pintner, R., and Forlano, G. "A Comparison of Methods of Item Selection for a Personality Test," *Journal of Applied Psychology,* **21**: 643–52, 1937.

64. Swineford, F. "Validity of Test Items," *Journal of Educational Psychology,* **27**: 68–78, 1936.

65. ———. "Biserial r Versus Pearson r as Measures of Test-Item Validity," *Journal of Educational Psychology,* **27**: 471–72, 1936.

66. Wesman, A. G. "Effect of Speed on Item-Test Correlation Coefficients," *Educational and Psychological Measurement,* **9**: 51–57, 1949.

67. Wherry, R. J., and Gaylord, R. H. "The Concept of Test and Item Reliability in Relation to Factor Pattern," *Psychometrika,* **8**: 247–64, 1943.

68. Zubin, J. "Note on a Transformation Function for Proportions and Percentages," *Journal of Applied Psychology,* **19**: 213–20, 1935.

Item Difficulty

69. Brogden, H. E. "Variation in Test Validity with Variation in the Distribution of Item Difficulties, Number of Items, and Degree of Their Intercorrelation," *Psychometrika,* **11**: 197–214, 1946.

70. CARROLL, J. B. "The Effect of Difficulty and Chance Success on Correlations Between Items or Between Tests," *Psychometrika,* **10**: 1–19, 1945.
71. CHEN, L. "The Correction Formula for Matching Tests," *Journal of Educational Psychology,* **35**: 565–66, 1944.
72. DAVIS, F. B. "The Selection of Test Items According to Difficulty," *American Psychologist,* **4**: 243, 1949. (Abstract.)
73. FERGUSON, G. A. "On the Theory of Test Discrimination," *Psychometrika,* **14**: 61–68, 1949.
74. GUILFORD, J. P. "The Determination of Item Difficulty When Chance Success Is a Factor," *Psychometrika,* **1**: 259–64, 1936.
75. GULLIKSEN, H. O. "The Relation of Item Difficulty and Inter-Item Correlation to Test Variance and Reliability," *Psychometrika,* **10**: 79–92, 1945.
76. HARRIS, C. W. "Prediction of the Difficulty Index of Objective-Type Spelling Items," *Educational and Psychological Measurement,* **7**: 319–25, 1947.
77. HORST, A. P. "The Chance Element in the Multiple-Choice Test Item," *Journal of General Psychology,* **6**: 209–11, 1932.
78. ———. "The Difficulty of a Multiple-Choice Test Item," *Journal of Educational Psychology,* **24**: 229–32, 1933.
79. JACKSON, R. W. B., and FERGUSON, G. A. "A Plea for a Functional Approach to Test Construction," *Educational and Psychological Measurement,* **3**: 23–28, 1943.
80. LAWLEY, D. N. "On Problems Connected with Item Selection and Test Construction," *Proceedings of the Royal Society of Edinburgh,* **61**: Section A, Part III 273–87, 1942–43.
81. ODELL, C. W. "The Scoring of Continuity or Rearrangement Tests," *Journal of Educational Psychology,* **35**: 352–56, 1944.
82. RICHARDSON, M. W. "The Relation Between the Difficulty and the Differential Validity of a Test," *Psychometrika,* **1**: 33–49, 1936.
83. SYMONDS, P. M. "Factors Influencing Test Reliability," *Journal of Educational Psychology,* **19**: 73–87, 1928.
84. THURSTONE, T. G. "The Difficulty of a Test and Its Diagnostic Value," *Journal of Educational Psychology,* **23**: 335–43, 1932.
85. WALKER, D. A. "Answer-Pattern and Score-Scatter in Tests and Examinations," *British Journal of Psychology,* **22**: 73–86, 1931; **26**: 301–8, 1936; **30**: 248–60, 1940.

THE COMBINATION OF TEST ITEMS

86. ADKINS, D. C., and TOOPS, H. A., "Simplified Formulas for Item Selection and Construction," *Psychometrika,* **2**: 165–71, 1937.
87. BROGDEN, H. E. "An Approach to the Problem of Differential Prediction," *Psychometrika,* **11**: 139–54, 1946.
88. FLANAGAN, J. C. "A Short Method for Selecting the Best Combination of Test Items for a Particular Purpose," *Psychological Bulletin,* **33**: 603–4, 1936. (Abstract.)
89. GULLIKSEN, H. O. *Selection of Test Items by Correlation With an External Criterion, as Applied to a Mechanical Comprehension Test.* (Office of Scientific Research and Development, Report No. 13319.) Washington: Department of Commerce. 1946.
90. HORST, A. P. "Determination of Optimal Test Length to Maximize the Multiple Correlation," *Psychometrika,* **14**: 79–88, 1949.
91. ———. "The Economical Collection of Data for Test Validation," *Journal of Experimental Education,* **2**: 250–53, 1934.
92. ———. "Item Analysis by the Method of Successive Residuals," *Journal of Experimental Education,* **2**: 254–63, 1934.
93. ———. "Item Selection by Means of a Maximizing Function," *Psychometrika,* **1**: 229–44, 1936.
94. LONG, W. F., and BURR, I. W. "Development of a Method of Increasing the Utility of Multiple Correlations by Considering Both Testing Time and Test Validity," *Psychometrika,* **14**: 137–61, 1949.

95. OWENS, W. A. "An Empirical Study of the Relationship Between Item Validity and Internal Consistency," *Educational and Psychological Measurement,* 7: 281–88, 1947.
96. RICHARDSON, M. W., and ADKINS, D. C. "A Rapid Method of Selecting Test Items," *Journal of Educational Psychology,* 29: 547–52, 1938.
97. TOOPS, H. A. "The L-Method," *Psychometrika,* 6: 249–66, 1941.

STATISTICAL TABLES AND REFERENCES

98. CHESIRE, L.; SAFFIR, M.; and THURSTONE, L. L. *Computing Diagrams for the Tetrachoric Correlation Coefficient.* Chicago: University of Chicago Bookstore, 1933.
99. FISHER, R. A. *Statistical Methods for Research Workers.* London: Oliver & Boyd, 1938.
100. ———, and YATES, F. *Statistical Tables for Biological, Agricultural, and Medical Research.* London: Oliver & Boyd, 1938.
101. KELLEY, T. L. *Fundamentals of Statistics.* Cambridge: Harvard University Press, 1947.
102. ———. *The Kelley Statistical Tables.* Cambridge: Harvard University Press, 1948.
103. ———. *Statistical Method.* New York: Macmillan Co., 1924.
104. MCNEMAR, Q. *Psychological Statistics.* New York: Wiley, 1949.
105. PEARSON, K. (ed.) *Tables for Statisticians and Biometricians, Part I.* London: Biometric Laboratory, University College, 1914.
106. ——— (ed.). *Tables for Statisticians and Biometricians, Part II.* London: Biometric Laboratory, University College, 1931.

10. Administering and Scoring the Objective Test

By ARTHUR E. TRAXLER
Educational Records Bureau

COLLABORATORS: Edward E. Cureton, *University of Tennessee;* Paul L. Dressel, *Michigan State College;* Herbert A. Toops, *Ohio State University*

THE REALIZATION OF THE POTENTIAL VALUES OF EDUCATIONAL measurement depends largely upon the understanding, accuracy, and competence with which tests are administered and used by the multitude of nonspecialists in measurement responsible for the teaching and guidance of the pupils in our schools. If test specialists are to avoid having their efforts to construct precise, well-standardized tests negated, they must take every precaution to safeguard the application and use of the tests. If school administrators are to find the results of tests worth the time and expense, great care must be taken to insure that the accuracy of the scores is not vitiated through misunderstanding or carelessness in the administration and scoring of the tests.

The Importance of the "Mechanical" Procedures

VALID RESULTS DEPEND UPON ACCURATE ADMINISTRATION AND SCORING

In view of their crucial importance in the whole chain of events from the conception of the test to the use of the scores in conferences with individuals, it seems highly unfortunate that the giving and scoring of tests are frequently treated very casually by both the authors and the users of tests. Test specialists have been dilatory in providing research data on the many debatable points relative to test administration and scoring, and, in general, test makers have not applied the same care and zeal to the writing of directions for administering tests that they have applied to item validation and other technical aspects of test construction. The reasons for special points in directions are seldom given, and needful precautions are seldom stated. Even when the directions have been carefully formulated, too often they have not been observed faithfully by test users.

The point should be stressed that tests are standardized on the basis of a particular set of directions for administering and scoring. Comparisons with norms are valid only when exactly the same procedure is used in administering and scoring the tests locally that was employed when the norms were established. It is likewise true that full comparability of scores from school to school within a system or a common testing program can be achieved only if all schools give the tests in the same way and have a maximum degree of scoring accuracy, or at least an equivalent bias in scoring errors. More important, still, valid comparisons from test to test—that is, the diagnostic values of tests for the individual—depend upon careful administration and accurate scoring of all the tests.

Inadequate Administration and Scoring May Vitiate Tests Results

There are numerous ways in which administration and scoring errors may impair test scores. Some types of improper administration or scoring cause bias in the results for entire groups and render intergroup comparisons of little value. Other kinds affect only certain individuals in varying degrees and thus lower the validity of the test results for guidance, counseling, and self-insight. Some of the more important inadequacies and errors in the giving and scoring of tests will be mentioned briefly. Most of them will be treated in greater detail later in the chapter.

Administration.—Probably the greatest single source of error in test administration is incorrect *timing* of tests that involve a time limit. Failure to read the timing directions correctly, lack of understanding on the part of inexperienced examiners of the need for precise timing, carelessness, faulty timepieces, and unintelligent scheduling so that it is impossible to allow enough time for a given test and still maintain the schedule—all these are potential sources of large errors which affect the scores of all individuals in the group. Test makers and publishers can help to prevent timing errors by emphasizing the need for precise timing, by outlining for the examiner a very definite time schedule from the very beginning of the test to its end, including rest periods, by showing the time in boldface print, and by summarizing the time limits in a time-administration table as well as having them interspersed at the proper places in the verbal directions.

Having time limits long enough that a time-limit test virtually is a work-limit test for a majority of the examinees serves to reduce the ill effects of bad test administration. This can be done, of course, only in those tests where the speed of response is not itself the object of the test. This procedure evades the chief fault of the work-limit method; namely, that it does

not fit well with fixed time schedules of the school. It has its disadvantage in discipline and motivation.

When a battery of tests is administered, directions for later tests can often be greatly shortened. It is often of value in such cases to write special directions for all tests. Those for the first are usually identical with the original ones; those for later tests are shortened as is appropriate to eliminate duplication. Failure to eliminate duplication makes it possible for some examinees to start a later test while the directions are being read, and results in boredom for the examinees and waste of time for all concerned.

A second source of large error in administration is *lack of clarity in the directions* to the examinees. Failure to make the directions clear to the examinees may result from an unusually complicated testing situation which is not explained in sufficient detail in the manual, blind spots on the part of the test maker in preparing the directions, confusion of the examinee due to inadequate preparation, slovenly reading of the directions by the examiner, use of a vocabulary beyond the level of the examinees, inattention of individual examinees, and too many directions or directions too far removed from their application. Confusion of directions is, in some respects, more serious than mistakes in timing. Errors in timing in effect either add or subtract a constant amount of time (but not of performance) for all pupils in the group, except those who would finish the test anyhow, and approximate adjustments in the scores can sometimes be made. Unclear directions, on the other hand, may lower the scores of either the entire group or of certain pupils in varying amounts. The test manual should contain suggestions to examiners concerning procedures in presenting the directions and in discussing examples, and specific recommendations concerning the extent to which informal supplementary explanation (if any) may properly be given to the pupils.

A third source of error in the administration, as well as in the scoring, of objective tests is failure to make clear to pupils what they are expected to do about *guessing,* or failure to indicate how, if at all, the scores are to be corrected for guessing. Although the problem created by guessing bulks large in test theory and research, it probably is, from a practical standpoint, a less important source of error than the first two mentioned. Nevertheless, this problem is important enough to warrant considerable attention later in this chapter.

A fourth source of error in test administration is variation in the *physical conditions* under which tests are administered. This area has been almost entirely neglected by test specialists, but there is reason to believe that sometimes it has an important bearing on test results. Such factors are more

likely to have an adverse effect on time-limit tests than on work-limit ones.

Schools vary widely in their facilities for the administration of tests to groups. In the matter of writing space alone, some schools use large, comfortable tables, others use desks, others armchairs, and still others give their tests in the auditorium with each pupil writing on a portable beaverboard "desk," or even on his lap. It is not reasonable to expect fully comparable results under such varying conditions. The current tendency to use separate answer sheets further complicates the problem created by differences in physical equipment. The examinee needs more desk space, horizontally, when employing separate answer sheets than otherwise. Test makers would do well to introduce into their manuals of directions a section on the environmental conditions under which the test may properly be given. They also have an obligation, incidentally, to see that the data for norms are obtained under defined environmental conditions and to report to users just what those conditions are.

Other sources of error in the giving of tests are too little or too much stress on *motivation* and failure to control opportunities for chance or purposeful *copying*. The problems in these two areas are largely those of the local examiner, although the test author can aid the examiner materially by means of appropriate suggestions in the test manual.

Scoring.—Scorers' tests administered to applicants for scoring positions by the Educational Records Bureau and other test service organizations reveal extremely wide differences in aptitude for this kind of work as measured in terms of accuracy and speed. If scoring is to be done well, individuals need to be carefully selected, trained, supervised, and checked upon systemically. Most significant scoring errors are caused by assignment of this work to individuals who are not accurate in clerical routine, by ineffective training, lack of adequate directions concerning procedure, and failure to see that all processes where large errors could occur are invariably checked by a second person.

On importance of checking.—Only a comparatively few people, even teachers, can add as many as three subtest scores correctly without making substantial errors occasionally, so this process should always be checked. Only a few know how to alphabetize, so unless this function is done by an expert, an examinee who has taken a half-dozen tests will show up with some missing when the test scores are assembled for a profile. In large test projects, the alphabetization of the test results becomes increasingly cumbersome. The process will be facilitated by imprinting on the answer medium five blocks for the printing-in by the examinee, in capital letters, of five letters: the initial letter of his last name, the second and third letters of his

last name, and the initial letters of his first and middle names. Thus, George Robert Johnson would respond |J|O|H|G|R|. Such matters save hundreds of hours of labor on large test projects. The principle is that a small amount of time (free) on the part of a large number of examinees will save a large amount of time of a (paid) clerical force. This principle has wide applicability. If examinees could reliably score their own or each other's papers in a few minutes time, no scoring device in which the papers have to be fed to the machine individually could compete with it.

As will be indicated in greater detail later, test makers can reduce the number of errors and increase scoring speed by the arrangement of the material on the page, by careful planning of the scoring keys, by preparation of separate answer sheets specially designed to facilitate scoring, and sometimes by overprinting of the key on the answer sheets by a second impression after the tests have been administered.

Need for Meticulous Attention to Details

There is a curious fallacy in the contemporary attitude toward the administration and scoring of group tests as compared with individual tests. It is doubtful if any educational institution would trust the administration and scoring of the individual Stanford-Binet Scale to anyone other than a trained psychometrician, but there is a rather general impression among school people that almost anyone can administer and score a group test if only he has a manual at hand. As a matter of fact, however, the administration of a group test to a class of pupils in one sense is a far more exacting procedure than the giving of an individual test to one pupil. The routine is more rigid and the penalty for error is multiplied by the number of individuals in the group. In the administration of an individual test a certain amount of leeway may be allowed in order to create a situation favorable for the eliciting of responses from that particular individual, but when testing a group, complete fidelity to all details of the prescribed procedure is imperative if the examiner is to avoid wasting the time of many individuals and if the results are to have the same meaning for all.

A high level of success in either the administration or scoring of tests is conditioned largely by a willingness on the part of the examiner to attend and by a habit of attending religiously to a host of details, any one of which may seem of little importance. The test author needs to outline every detail clearly, completely, and with convincing emphasis upon its importance. The examiner must understand that each detail is an integral and indispensable part of the standardization process and necessary for adequate comparisons of local groups with test norms.

The Administration of Tests

As suggested in the preceding section, the accurate and efficient giving of a test requires close cooperation between the test author and the examiner. It is desirable to review in considerable detail the responsibilities of both these functionaries with regard to test administration.

Administration from the Viewpoint of the Test Maker

Although it is occasionally necessary for the examiner to make minor adjustments to fit the local situation, the conditions of test administration should, as far as possible, be under the control of the test author, especially if use is to be made of the norms. Comparability in the results from class to class, school to school, and test to test can be expected only when this principle is observed. It is, therefore, one of the major responsibilities of the test maker to decide and specify clearly all conditions of administration. The preparation of directions for giving the test is universally accepted as a phase of test construction, but too often it is regarded as a minor aspect to be handled hastily just as the test is about to go to press. The formulation of the instructions, including the application of research in reaching decisions, should be regarded as a professional task equally important with item writing and validation.

Among the main decisions which a test author needs to make before he writes the directions for administration are those concerned with time limits, motivation of the subjects, and control of guessing.

Setting of time limits

In measurement dealing with "higher thought processes," counselors, school administrators, and the subjects themselves usually are primarily interested in determining *level* rather than *speed* of work. Speed is admittedly of some individual importance, particularly when the rate of work is exceptionally rapid or exceedingly slow in connection with certain skills constantly in use, such as reading, alphabetization, and the like. It is also decidedly important in some mechanical jobs. Since most life situations, however, ordinarily do not call for completion of a certain task within rigid time limits, success generally is conditioned by level more than by speed. Two individuals with equal level of achievement may be deemed equally successful on a job even though one is twice as fast as the other, if the second one is able and willing to spend twice as much time on the task. Theoretically, most educational tests might well be untimed if efficient use of testing time were not important and if it were not necessary for testing programs to conform to school time-scheduling.

The school day is usually divided into class periods of forty to sixty minutes' duration. It is no doubt desirable for schools to plan special schedules to handle their testing programs, but many schools do not find it feasible to do this. Consequently, there is widespread demand on the part of schools for tests that will conform to the customary class periods. Most test authors accede to this demand and set their tests for approximately forty minutes, or multiples thereof. The result is that many achievement tests are entirely too short to permit adequate reliability of performance, to say nothing of validity of response.

In order to take account of class periods, a test maker may either use more material than most pupils can cover in the period and set up a rigid time schedule, or he may make his test primarily one of level by using so small an amount of material that all, or practically all, pupils can complete it within the designated period. The second procedure is used sparingly, for it tends to lower the statistical reliability of the test, and, moreover, it frequently is frowned upon by school people since they feel that it complicates the discipline problem during test administration and encourages idleness and time wasting on the part of the faster pupils. The upshot is that most tests designed for school use are timed tests.

On objections to time-limit tests.—If the disciplinary problem created by having pupils finish at different times can be surmounted, many of the difficulties of time limits are avoided by employing work-limit tests. The latter afford generally better motivation. Furthermore, since all items are attempted by everyone, the test responses may be analyzed to pick out those items most worthy of perpetuation in subsequent editions of the tests—assuming that a criterion is, or can, be made available. If this is done, the individual items of a test should improve in statistical merit in subsequent editions. This is difficult to achieve in time-limit tests. Time-limit tests, furthermore, are prone to yield spuriously high reliability coefficients when they are computed by split-halves or internal consistency methods.

Relationship of test length to objectives of measurement.—Theoretically, the length of a test and the time limit for it should be determined in the light of the purpose for which the test is given. It is well known that tests used to evaluate *groups* may be very short, for reliable differences among groups may be obtained on the basis of a small number of questions. If the purpose is *over-all appraisal of individuals,* in terms of total score, the test must be longer, but for few aspects of school aptitude or achievement does it need to be longer than one class period in order to be satisfactorily reliable. If *individual diagnosis* is desired, a total of three or four class periods may be called for in order to use enough test items to bring about

sufficient reliability in the several diagnostic scores. Considering its potential importance to the individual examinee, he could profitably devote, if need be, many times the number of hours he now devotes to test taking.

As a matter of fact, it is seldom feasible, aside from specific experimental situations, to separate objectives in this way. Schools tend to use a single test for a variety of objectives. They wish to make interschool comparisons, to evaluate class groups, to provide data for the counseling of individuals, and to diagnose pupil strengths and weaknesses, all on the basis of tests requiring exactly one class period. The test maker usually compromises by aiming first of all at obtaining valid and reliable total scores of individuals while at the same time providing an opportunity for limited diagnosis through setting up part scores on three or four of the more important areas covered by the test.

Need for experimental determination of optimum time limits.—There is a definite need for test makers to determine on the basis of research the amount of time to be allowed for different types of material. Theoretically, the problem of setting time limits for educational achievement tests involves finding the answer to either of two questions: (1) "For a given amount of test material, what amount of *time per item* will result in the *maximum validity per unit of time?*" or, (2) if the test is to fit a certain predetermined time interval, "What number of items will yield the maximum validity in the designated time?"

One section of a study by Cook (*11*) illustrates the application of this type of experimentation to the measurement of spelling ability. Cook defined the optimum time for the administration of the test as "that time at which further increase in the validity of the obtained scores can better be secured through the addition of more material with a proportionate increase in time than by permitting more time on the same material." He then applied this definition to various arrangements of a 50-word spelling test at the eighth-grade level, using as a criterion a 150-word list dictation spelling test. His results indicated optimum rates in the testing of eighth-grade pupils, such as ten words a minute for a right-wrong spelling test, five words a minute for a four-response spelling test, and about four words a minute for a word in a sentence-recall spelling test.

This technique assumes that an external criterion of validity is available. The criterion may be concerned exclusively with power or level to the complete exclusion of speed, but even in this situation the "best" time limit for the experimental test is not necessarily an indefinite or even a liberal time limit. For example, the total score on a 100-item experimental test might correlate .80 with a given criterion, if the administration time were liberal

enough to allow all pupils to finish, say two hours, and might show a lower correlation with the criterion if administered in any shorter time limit, in which only part of the pupils finish. This finding would not mean that the best results will be obtained by administering this 100-item test in two hours. It is possible that the great majority of the examinees are able to complete this test in one hour, and that the time of most of them is wasted, with no improvement in the validity of their scores, in order to obtain a slight increase in validity in the scores of just two or three slow workers. It might be better, therefore, to improve the validity of the majority of the scores by increasing the length of the test (but not the time), even at some sacrifice in the validity of the scores of the few slow examinees. That is, one might secure a more valid score for the typical examinee by administering *more* than 100 items in the two hours. Experimentation might show, for example, that 164 items administered in two hours might yield a total score that would correlate .82 with the criterion. More items would yield a lower correlation because too much emphasis is then placed on speed; fewer items would yield a lower correlation because of the more limited sampling of content. It is true that the 164 items might yield a still more valid score in a longer time period, but they would, nevertheless, represent the most efficient way to use the *two hours*.

If speed is not a factor in the criterion at all, the total score on a given amount of test material will yield the maximum total validity if administered under work-limit conditions, but not the maximum validity *per unit of time*. Thus, even though speed is not a factor in the criterion at all, one might make most efficient use of whatever amount of time is taken, and improve the validity of the majority of scores, if a time limit is employed which will not permit all pupils to finish. This principle should be recognized in achievement test construction in general, even though external criteria are not available to make possible the experimental determination of optimum time limits.

In actual practice, of course, external criteria of validity are rarely available in the construction of educational achievement tests. It is usually necessary to compromise between what is sometimes regarded as the "ideal" work-limit conditions and various practical considerations. It is known that pupils differ greatly in speed and that if a test is administered as a work-limit test in which even the very slowest workers are allowed to finish, either a considerable amount of time will be wasted for most of the pupils and a disciplinary problem will be created in attempting to keep all pupils in their seats until all have finished, or irrelevant distractions will be introduced for pupils still working if the faster pupils are permitted to leave

whenever they finish the test. Accordingly, test authors usually set the time limits so that between 80 and 90 percent of the pupils can consider or attempt all of the items and use various devices (such as "cushion" items of high difficulty at the end of the test and interspersed directions throughout the test period) to keep the fast pupils occupied and to keep the slow pupils moving along more rapidly. Actually, higher validity per unit of time is probably secured on this basis than would be obtained if the same test were administered under work-limit conditions.

There is a fertile and almost untouched field for research on the optimum rates of administration of tests in many subject fields and at different grade levels.

Relation of timing directions to the degree of independence in the variables sampled by the parts of a test.—Two general procedures are used in timing the different parts of a test. One procedure is to time each part entirely independently through having all pupils start a part at exactly the same time and then wait for the signal, if they finish earlier, to go on to the next part. The other is to instruct the students in the beginning that if they finish a part before time is called they are to go on to the next part, and then to say at the expiration of the time for a given part, "If you have not finished Part I, stop working on that part and go on to Part II," and so forth.

The theory back of that kind of timing is that, other things being equal, the more items attempted, the higher the reliability and, indirectly, the higher the validity as well. A pupil who is rapid on one type of material may be slow on another type. If the test booklet, as a whole, contains enough material to keep most pupils working continuously throughout the entire period, it is reasoned that the reliability and validity of the total scores will be higher if the pupils do not spend time waiting for signals to proceed when they might be using that time in the solution of questions. Moreover, as indicated earlier, level rather than speed tends to be stressed when the subjects are allowed to use all the time in the testing period to full advantage. With the time limit about right, the directions serve only to keep the very slowest from taking an interminable amount of time on individual items.

Although the logic of that point is obvious as far as the total score is concerned, it appears that decisions concerning exclusive or overlapping timing of the parts should be made on the basis of whether or not separate part scores are to be obtained and of whether the functions measured by the parts are highly correlated or relatively independent. The timing of the Cooperative Reading Comprehension Test is a case in point. This test has two parts—vocabulary and reading comprehension. It yields scores for vocabu-

lary, speed of comprehension, and level of comprehension. Fifteen minutes are allowed for the vocabulary part and twenty-five minutes for the reading comprehension part. The examinees are instructed to go on to the reading comprehension part as soon as they have finished vocabulary even though the time for that part has not expired. Thus, the speed-of-comprehension score (but not the level score) represents a composite of rate of responding to vocabulary items and rate of reading paragraphs and answering questions. This way of timing is defensible only if it can be shown that these functions are so highly correlated that the speed-of-comprehension Scaled Scores of the pupils under this overlapping time arrangement are virtually equivalent to, or better than, what they would be if the time limits were independent. Research on such matters is virtually nonexistent.

Correlations among speed, level, and power tests.—The question of the advisability of using time limits with tests designed to measure scholastic aptitude and achievement should be considered in the light of the correlations between timed and untimed tests. It is known that when speed and level are measured independently the correlation between the resulting scores is comparatively low. Blommers and Lindquist (*6*) reported that the within-grade, within-school correlation of rate of reading comprehension with reading comprehension from which the influence of speed was eliminated was approximately .30. Tinker (*56*) found that speed and level scores on the Revised Minnesota Paper Form Board Test varied independently and that about 75 percent of the variance of the time-limit score, which he called the "power" score was accounted for by speed and level. Likewise, Davidson and Carroll (*17*) in a factor analysis of a number of group mental tests, administered to college students, found that speed was linearly independent of level and that time-limit scores could be represented as factorially complex measures having heavy loadings of both speed and level. They concluded that because of their factorial complexity, time-limit scores should be used with considerable caution in the prediction of criteria.

Other studies, however, have shown that when timed and untimed scores (liberal time limits) are obtained from the same test at the same sitting the correlations are comparatively high. For example, the correlation between speed of comprehension and level of comprehension on the Cooperative Literary Comprehension Test was found in one study (*64*) to be .917. While theoretically it would usually be preferable to administer most achievement tests without time limits if ample time were available for testing, the correlation between untimed, or "level," scores, and liberally timed, or "speed," scores usually is high enough to warrant the administration of tests under liberally timed conditions.

Eliminating the influence of time on certain tests.—The most common

way of reducing or eliminating the influence of time on tests is to set the time limits so liberally that all, or nearly all, pupils are able to consider or attempt all the items in the test. It is apparently the intention of many achievement test constructors at present to allow from 80 to 90 percent of the pupils in a typical group to finish a section before instructing the group to turn to the next section.

It is possible, however, to reduce the effect of the time factor in some tests through ingenuity in test construction without either cutting the amount of material in the test or allowing more time. A good illustration of one procedure for obtaining a level score in a timed test is furnished by the Cooperative Reading Comprehension Test. The paragraph reading part of this test contains 90 items arranged in three cycles of 30 items each. Since the three cycles are of comparable difficulty, each cycle is a shortened form of the entire test. The level of comprehension score is based on the average number of correct responses in the cycle *completed* by each subject, as contrasted with the speed-of-comprehension score, which is found from the number of correct responses in the time limit. Even the slowest individuals nearly always finish the first cycle and thus obtain level scores which are not dependent upon the time allowed for the test. This technique could appropriately be applied to other kinds of tests. It is no doubt better adapted to the measurement of scholastic aptitude, reading ability, and skills than to achievement testing in the content fields.

Devices to insure careful observation of time limits.—Several suggestions concerning the formulation of directions so as to facilitate observation of time limits were made in an earlier section of this chapter. A further device to encourage examiners to give careful attention to the timing of tests is a room examiner's report sheet. This sheet summarizes the allotted time for the different tests and provides blank spaces where the examiner may record the starting time and the finishing time actually employed. The sheet should be filled out and signed by the room examiner and then turned in to the head examiner of the local school system, who, in turn, should forward it to the central agency if the school is participating in a state-wide or nation-wide testing program. To illustrate this procedure a report sheet used in one of the testing programs carried on by the member schools of the Educational Records Bureau is shown in Figure 6 and a schedule employed in the National Teacher Examinations is exhibited in Figure 7.

Considerable impetus can be given to accurate timing through the provision of better timepieces. A survey of timing procedures in the administration of tests by member schools of the Educational Records Bureau indicated that nearly half of the examiners used ordinary watches and that from 20 to 40 percent of them did not check their watches for accuracy before timing the tests.

EXAMINERS' REPORT SHEET
for
AMERICAN COUNCIL PSYCHOLOGICAL EXAMINATION, IOWA SILENT READING TEST,
AND PROGRESSIVE ACHIEVEMENT TEST

Name of test ______________________________

No. of pupils ________ No. of proctors ________ Were pupils seated in alternate seats? ________ Type of desk used by pupils ______________________________

Was ordinary or stop watch used? ______________________________

Was manual of directions studied carefully before starting test? ________________

The number of minutes required for the various parts of each test is shown in the table below. Immediately after starting each part record the starting time and the finishing time in columns 4 and 5. (Examiners using stop watches need not fill in columns 4 and 5.)

A.C.Psych. Minutes Required (1936 edition)	Iowa Reading Adv. Minutes Required (1927 edition)	Progressive Achievement Int. or Adv. Min. Required	Start. Time	Finish. Time	Prog.Ach. Test	Start. Time	Finish. Time
		Int. Adv.			Int. Adv.		
I 10	I A 5	I A 3 4			IVF 10 9		
II 20	B 5	B 3 4			G 12 10		
III 13	II A 2	C 3 4			H 12 10		
IV 10	B 1½	D 3 4			V A 3 3		
V 7	C 1½	II E 8 8			B 2 2		
	D 2	F 5 6			C 4 5		
	III A 3	G 25 20			D 5 10		
	B 3	III A 4 5			E - -	--	--
	IV 5	B 5 5			F - -	--	--
	V A 2	C 5 5					
	B 3	D 16 15					
	VI 2	IV E 10 9					

Irregularities (Describe in detail; use back of sheet, if necessary. Include names of any pupils leaving room during examination and amount of time out.)

____

____

General comments. (These will be especially welcome. Use back of sheet if more space is needed.)

Examiner ______________________

School ______________________

FIG. 6.—Example of test report sheet to be signed by room examiner. Used by member schools of the Educational Records Bureau.

SCHEDULE OF EXAMINATIONS
AND
DIRECTIONS TO BE READ TO CANDIDATES DURING THE TESTING

February 9, February 16, 1946

Session / Test / Part	Time to Allow: (Minutes)	Time Record* Start Hr.	Start Min.	Stop Hr.	Stop Min.	Direction To Give Beginning	End
First Session (Feb. 9)						1, 2	
Report Cards	30					3	
Verbal Comprehension							
Part I (Read.)	20					4	5
Part II (Vocab.)	10						6
Non-verbal Reasoning							
Part I (Fig. Anal.)	10						7
Part II (Pat. Reas.)	10						8
Part III (No. Series)	10						9
Recess	5						10
Professional Information							
Part I (Ed. Soc. Pol.)	25						11
Part II (Ed. Psy.)	25						11
Part III (Guid. & Anal.)	25						11
Part IV (Prin. & Meth.)	25						12
Second Session (Feb. 9)						13, 2, 15	
English Expression							
Part I (Mech.)	15					14	16
Part II (Effect.)	15						17
General Culture							
Part I							
Sect. 1	20						11
Sect. 2	20						11
Sect. 3	20						9
Recess	5						10
General Culture							
Part II							
Sect. 1	20						11
Sect. 2	20						11
Part III							
Sect. 1	25						11
Sect. 2	25						18
Third Session (Feb. 16)						19, 2	
Report Cards	30					20	
Option I							
Part I	30					21, 24	11
Part II							
Sect. 1**	15						23
Sect. 2	15						11
Part III							
Sect. 1**	15						23
Sect. 2	15						23
Sect. 3	10						25
Recess	5						22, 24
Option II							
Part I	30						11
Part II							
Sect. 1**	15						23
Sect. 2	15						11
Part III							
Sect. 1**	15						23
Sect. 2	15						23
Sect. 3	10						26

*To be filled in by Examiner.
**If Education in Elementary School option is not being administered, timing by sections of Parts II and III is not required. Total time Part II, 30 minutes; Part III, 40 minutes.

This Schedule and Time Record is to be signed by Room Examiner and submitted to Local Examiner for forwarding at end of last examining session.

____________________ Room Examiner

Local Examiner ____________________

Center ____________________

FIG. 7.—Example of schedule and time record to be signed by room examiner administering National Teacher Examinations.

Because of the expense involved, few schools would find it feasible to provide each staff member serving as an examiner in a group testing program with a stop watch, nor is a watch graduated into fifths of a second needed in timing group tests. Some stop watches require rewinding after about a half hour's use; and the time in the last ten minutes is likely to be somewhat in error. A very satisfactory substitute for, or in some respects a distinct improvement over, a stop watch is an interval timer, which is simply a miniature alarm clock which rings a bell at the end of any set interval of time. Another good illustration is a sports timer, which is simply an ordinary, inexpensive watch with a start-and-stop device. Using this device, the examiner can stop the second hand at zero, set the minute and hour hands at 12:00, start the watch when he says "begin," and compute all time for the test from twelve o'clock. If it is necessary to take time out for directions between parts, the examiner needs merely to stop the watch while he reads the directions.

Experience has shown that this type of watch is especially helpful in reducing the examiner's tension and enabling him to do an accurate job. Sports timers and interval timers are easily within the budgets of nearly all schools. They do not get out of adjustment so easily as a stop watch, and with care they will last indefinitely. If a school's measurement department were equipped with a sufficient supply of these timers, they could be used for all group testing, and their accuracy could be checked under the supervision of the head examiner before they were handed out to the room examiners. It is believed that the use of this simple and inexpensive equipment would significantly increase the accuracy of measurement in schoolwide testing programs.

If employing an accurate, ordinary watch of standard make, one may say "go" as the second hand reaches 60. One minute later he should test the watch to see that it now reads exactly 12:01 as the second hand reaches 60. Old watches, even of standard make, often have so much lost motion in the gears that one may find it impossible after a time to be sure whether the time is, say, three minutes or four minutes. Such watches should never be used.

Even with good timers, serious difficulty is sometimes encountered because the timepiece fails during a testing period. To be safe, one ought to be able to refer in an emergency to a wall clock or a pocket watch in addition to a special timer.

Motivation

One of the most intangible of the variables affecting test performance is motivation. Since his directions help to condition motivation, the test maker should give careful attention to this problem.

Influence of motivation on test scores.—It is reasonable to believe that test scores are sometimes significantly influenced by motivation.

Glick, in 1925, showed that by taking five forms of Army Alpha, interspersed with fifteen practice forms, the average college sophomore could double his Army Alpha score (*27*). Courtis, in 1932, by using a wider variety of incentives showed an even greater improvement (*12*). Such incentives, however, would never normally prevail in a testing situation.

Kirlin (*36*) carried on a controlled experiment in which Forms A and B of the Advanced Compass Survey Tests in Arithmetic were administered to 131 pupils in grades eight and nine. The two forms were first administered without special motivation, with half the pupils taking one form and the other half the other form. Three days later the forms were reversed and were given with every possible effort to motivate the pupils. In the preliminary instructions for the second testing, the pupils were told that their scores would help to determine their six weeks' grades, and cash prizes were offered to the pupils making the largest increases in scores over the first testing. Observation indicated that the pupils were keenly interested and greatly motivated. The results showed a statistically significant increase in mean scores, but on the average not enough to change a pupil's percentile rank in the entire group by more than a few points. It is probable, also, that practice effect accounted for part of this difference.

In view of the probable influence of practice and the fact that means of motivation were employed that were beyond those which would ordinarily be used in a school situation, the gains found by those who have studied the question of motivation may be discounted to some extent. The results indicate that motivation does affect test performance, but perhaps not to as great an extent as commonly supposed. This is another area for research that needs further exploration.

Non-test-motivated individuals are commonly believed to exist. Cases of such are occasionally produced. In the Army, the test-malingerer is always a potential or actual problem. Goldstein devised a scoring procedure for detecting malingerers among those taking tests in the Army and showed experimentally that this scoring key distinguished between malingerers and genuine failures in a large proportion of the cases (*28*). The theory of a potential index of test-taking motivation is simple:

1. Individuals are not able to gauge very accurately the difficulty of an item.

2. Even if they were able, they would not know, without an extensive statistical analysis (always wanting), how to regulate their score on easy items and on hard items of the same test.

3. Consequently, if the score on certain easy items is not statistically

comparable with the score on certain hard items of the same test, the examinee may be suspected of wrong test motivation.

4. Only with optimal motivation will the individual make comparable scores on two tests of unequal difficulty but otherwise comparable.

Even copying is potentially detectable by a greater generalization of the above notion. To remove such non-motivated or non-evenly-motivated individuals from our test populations would improve our item selection, our norms, our validity coefficients, and our statistical comparisons generally. It may not be too much to prophesy that shortly we shall not think of giving a test without deriving a score of test-taking motivation to tell us whether we shall take the test result at full value or only after a retest with more nearly optimal motivation prevailing. This is a problem for test authors and publishers and is one that they ought to consider seriously in future test construction. It is probably fair to say that, in general, modern test makers do a competent technical job but often do not take advantage of the opportunity to introduce innovations that would greatly assist the test user.

How much motivation is desirable?—The problem of motivation is undoubtedly complicated by the fact that individuals react to a motivating stimulus in different ways and that even before the stimulus is applied they are motivated in varying degrees by their individual reactions or "set" to the very idea of taking a test. Some pupils approach a test with complete lack of interest and considerable boredom and will not put forth their best efforts unless some outside motivating influence is present. Others become nervous and overexcited in a test situation and thus may find their performance blocked by their emotional state. The most desirable motivating conditions are those which enable the largest number of individuals to turn in the best performance without undue emotional stress.

Ways of standardizing or controlling motivation.—The personality and attitude of the examiner, of course, have a great deal to do with the motivation of the subjects. With a businesslike but not too severe manner, an alert and understanding examiner can arouse interest and, at the same time, relieve tension. He can, however, do a better job if he has the aid of brief printed directions in which the problem of motivation is recognized.

One kind of printed material useful in getting all individuals oriented to the tests and properly motivated for them is a general statement concerning the nature and purpose of the testing program which may be distributed to the pupils several days before the tests are given. A test-making or test-service agency probably should be prepared to furnish the schools it services with such a statement. The statement of "Test Information for Students" which the Educational Records Bureau distributes to its member schools is as follows:

TEST INFORMATION FOR STUDENTS

The purpose of this sheet is to inform you about the general nature of the Educational Records Bureau tests which you will take within the next few days. Two kinds of tests will be taken by practically all pupils. One of the tests is designed to determine your reading ability; the other is intended to measure your aptitude for school work. Your school may be one of those that will give still other tests to discover your knowledge of the subjects you have studied in school. None of the tests will be used to determine your grade or mark. The purpose in giving them is to inform your teachers about your ability and your needs so that they can provide the best possible learning conditions for you.

Each test contains a large number of questions calling for very brief answers, such as the writing of a word or two, the underlining of a word, or the selection of the correct answer from several suggested ones. Do not become discouraged if you find a large number of questions which you are unable to answer. You are not expected to answer all the questions. Some tests are used throughout several grades and, of course, the pupils in the lower grades are not expected to answer as many questions as those in the higher ones. It is practically impossible for even the most advanced students to obtain a perfect score.

It is advisable to answer some questions about which you are not entirely sure. If you think you know the answer, you should put it down even though you are not certain, but you should not guess wildly on questions concerning which you are totally ignorant. In some tests, in which a certain proportion of the wrong answers is subtracted from the correct answers, blind guessing may result in a large reduction in one's score.

There is no passing mark for these tests. The results will be expressed in percentiles which will show how you stand in comparison with other students who take the same tests in other schools.

Since time is such an important element, be sure to have at least two well-sharpened pencils and an eraser. It is not advisable to use ink because of the possibility of the pen running dry and the difficulty of erasing.

Some tests are to be given with special answer sheets in order that they may be scored by means of an electrical machine. If you take one of these tests, you will be given a practice test to acquaint you with the way in which the answer sheets should be marked. The important thing to remember is to use *only* the special pencil which will be furnished you and to make *heavy black marks.*

POINTS TO KEEP IN MIND WHEN TAKING THE TESTS

1. *Listen carefully* to all instructions given by the examiner.
2. You are not expected to answer *all* of the questions, but *answer as many as you can.*
3. Work as *rapidly* as you can, spending no time puzzling over difficult questions. Return to the hard questions if you have time after you have gone through the test.
4. Guess only if you can do so *intelligently.* Don't guess if you know nothing about the question.

5. Go prepared with *two pencils* and eraser. If a special pencil is given you, use it *only*.
6. Do not waste your spare time during the days on which the examinations are scheduled, but spend the time constructively as you would during any examination period.

The other kind of printed material useful in helping examiners deal with the problem of motivation is the printed instructions which the test maker includes in his manual for the examiner to read at the opening of the test period. These instructions are well illustrated by two paragraphs from the Manual of Directions for Iowa Tests of Educational Development:

> This morning you are going to begin taking a special series of tests. The purpose of these tests is to tell us something about your general educational development. We want to know what background of ideas you have in the social studies and in the natural sciences, how well you can read and interpret different kinds of materials, how readily you can solve arithmetical problems, etc. This information will help us to discover how well this high school as a whole is accomplishing some of the things it is supposed to be doing, and will show how it compares with other high schools in these respects. More important, what we learn about you from these tests will help your teachers to decide how to fit their teaching to your individual needs, and will also help us in advising you on your future educational plans. The test results can also be of real value to you in working out your own educational and vocational plans, and in making up your mind how best to distribute your effort in your school work.
>
> The test results can be very valuable in these ways, but only if each of you does his very best on all of the tests. If you do not make a sincere effort, the scores will tell a false and misleading story about your abilities, and your time spent in taking the tests will be worse than wasted. If you look upon these tests as a challenge—as an opportunity for you to show what you are really capable of doing—then you are bound to enjoy taking them. It is only if you do not try that the tests can become tiresome or uninteresting.

Guessing on objective tests

The question of what to do about chance successes and guessing has plagued test makers from the very beginning of objective testing. After more than three decades of experience with multiple-response tests, there is still not complete agreement among test specialists concerning this question.

Effect of guessing when score is number right.—All types of objective questions in which the subject is given a choice between two or more suggested responses, of course, involve elements of chance and guessing. The contributions of chance and guessing to the results of true-false and other two-response questions are, of course, potentially greater than the contributions of these same factors to the results of questions involving a

larger number of choices. Consequently, test makers have been more concerned with the effect of guessing on two-response tests than on tests involving a larger number of choices, but it is recognized that the problem continues to be present as the number of choices is increased, although in diminishing degrees.

The general effect of guessing, where the score is taken simply as the number right, is, of course, to raise the scores of the guessers. It is apparent that a pupil who knows nothing at all about a subject should be able by chance to answer correctly half of the questions in a true-false test on the subject. The more he knows about the subject, the less he will have to guess, but the theory is that if he guesses at all, he will, on the average, get half his guesses right and the other half wrong. By similar reasoning, on a three-response test, an individual has one chance of guessing right and two chances of guessing wrong; on a four-response test, one chance of being right and three chances of being wrong; on a five-response test, one chance of hitting upon the right answer by guessing and four chances of choosing the wrong answer, and so forth. The number right should, therefore, be discounted by $W/(n - 1)$. Hence, the corrected score, assuming that all incorrect responses are the result of guessing, is $S = R - W/(n - 1)$, where R is the number right, W the number wrong, and n the number of choices in each item.

If all pupils did an equal amount of guessing and if chance always operated according to theory, guessing would be of no importance in multiple-response tests unless one were interested in turning the scores into percentage grades. However, examinees vary greatly in their willingness to take a chance by guessing at questions about which they know nothing, and some will need to guess on fewer items than others. Thus, where the score is the number right, individuals will not have exactly the same relative standing on a test that they would have had if guessing had not been a variable in the scores.

There is a worth-while distinction to be made between "honest" and "dishonest" guessing. During the war many of the tests used in connection with the armed forces were scored on the basis of right answers. A common practice of men taking the tests who found they could not finish was to run rapidly through the remaining items and to mark some response without even reading the item. This is distinctly different from "honest" guessing based on hunches or half-knowledge. The correction for guessing is appropriate for the purely chance "dishonest" guessing, but no correction can be altogether appropriate for guessing based on some knowledge of the item. In everyday life, persons are continually guessing on the basis of partial knowledge, but they hope to guess right a fair share of the

time. If one who is taking a test can make intelligent guesses in excess of the correction for guessing, he is receiving only his just dues.

This point of view places the correction for guessing in a different light. One of the main purposes of the correction is to discourage dishonest guessing and to extract a suitable penalty for it.

Another very important consideration is that the proportion of examinees selecting any incorrect response depends upon its plausibility, and that the skillful item writer attempts to make each incorrect response as plausible as possible to the examiner who does not possess the desired knowledge or ability. The preparation of wrong responses, and hence of right responses also, to a multiple-choice item is thus dependent upon the skill of the item writer. In effect, the item writer attempts to make each wrong response so plausible that *every* examinee who does not possess the desired skill or ability will select a wrong response. In other words, the item writer's aim is to make all or nearly all *considered* guesses wrong guesses. If the item writer succeeded in this aim, and if all guesses were *considered* guesses, there would be no need for corrections for guessing. In such a situation, the standard scoring formula would seriously *overcorrect* for guessing. In actual practice, this aim of the item writer is never fully realized, but it is doubtless often sufficiently realized that the standard formula markedly *overcorrects*.

Still another aspect of this problem deserves serious consideration. As already noted, the correction for guessing formula assumes that *all* incorrect responses are the result of blind guessing, and that by subtracting $W/(n-1)$ from the number of rights one can get an estimate of the number of questions to which the examinee "really knows" the right answer. This logic seems fairly plausible as applied to a *right-answer* type of test (see page 196), but it is obviously inapplicable to a *best*-answer test. Items of the *best*-answer type require the student to distinguish between a number of responses all of which may be "right" to some degree, or none of which may be unqualifiedly correct. There is no "right" answer that can be "really known" by the examinee—*none* of the answers may ever have occurred to him before taking the test. The task for the examinee is one of comparison, and evaluation of all responses, not of recognition of a single *right* answer. For this kind of item, one can hardly discover how many "really knew" the right answers by subtracting a fraction of the choices of distracters.

Instructions to pupils concerning guessing.—There are two general procedures for dealing with the problem of guessing on objective tests. One is either to encourage guessing or to leave decisions concerning whether or not to guess to the individual idiosyncrasies of the examinees and then

to use a formula to attempt to correct for guessing as a part of the scoring process. The other is to try to control guessing either by instructing the examinees to guess or by telling them not to guess.

The theory of correction formulas and procedures of using such formulas will be discussed in the section on scoring. We are here concerned with the instructions which the test maker should prepare for the examiner to read to the pupils about guessing. An attempt may be made to eliminate guessing by instructing the examinees not to guess. If all individuals could follow uniformly an instruction of this kind, perhaps most of the problem of guessing would disappear. However, they cannot follow such instructions uniformly, since the meaning of the term "guessing" differs markedly from individual to individual. Instructions of this kind may serve only to magnify the effect of guessing upon the relative standing of the individuals in a group, for some individuals will observe the directions religiously, while others will disregard them.

From the standpoint of psychological and statistical theory, a strong case may be made out for instructing examinees to attempt every item in objective tests. Even if the scores are not corrected for guessing, it is felt that instructing pupils to try all items will reduce the differences in the scores due to guessing, since it will encourage the less venturesome individuals to do as much guessing as those who naturally have a greater tendency to take a chance. If a correction formula is applied, and if the pupils have followed instructions, the corrected scores will correlate perfectly with the uncorrected (rights) scores. It is true, of course, that there will still be some differences among the pupils in the extent to which they will follow instructions, but it is probable that, on the whole, there will be more uniformity in obedience to a positive instruction of this type than to a negative instruction not to guess.

However, there are two serious objections to this procedure of directing all examinees to mark all items. One is that it increases the error variance in the scores. If the examinee guesses on some items, his score is in part dependent upon his luck in those guesses, which would, of course, vary considerably from one examinee to another. If there is no guessing, this source of error is eliminated.

Another and perhaps more serious objection is concerned with the psychological effect of instruction to guess upon pupils and teachers. Regardless of the opinion of the psychologists and test specialists on this question, many school administrators and teachers object strenuously to the idea of directing pupils to attempt every item in a test. They say that among their students there is already too much carelessness, loose thinking, and guessing and that the educational implications of tests in which this tendency is

aided and abetted are decidedly bad. Some teachers feel that instead of encouraging guessing on objective tests, schools should take vigorous steps to discourage it. For example, Count Etoxinod (*24*) (pseudonym), after deploring the lack of respect for accuracy on the part of American students, proposed that the penalty for wrong answers on true-false tests be doubled in order to discourage guessing and to promote habits of careful thinking. Because of the apparent relationship of tests to educational practice, it is not probable that test specialists will ever be able to convince school people generally of the desirability of instructing pupils to guess when they take objective tests.

Under the circumstances, it seems probable that the most desirable statement concerning guessing that a test maker can include in his manual is a compromise between the guess-and-do-not-guess instructions. This kind of direction will carry more weight if the examinees know that the scores are to be corrected for guessing. For instance, the Educational Testing Service uses the following sentence on the cover page of the Cooperative Achievement Tests: "You may answer questions even when you are not perfectly sure your answers are correct, but you should avoid *wild* guessing, since wrong answers will result in a subtraction from the number of your correct answers."

One of the most practical reasons for correction for guessing is found in the fact that regardless of what instruction test makers and examiners give to the pupils, there may be systematic differences in test performance from school to school and from class to class because of differences in pretesting instructions given by individual teachers. Some teachers in the interest of promoting habits of careful thinking and accuracy will warn their pupils against guessing; others anxious to have their own classes make a good showing will urge their pupils to attempt all items; still others will explain to the pupils how, if they are able to eliminate certain choices in a multiple-response item and reduce the chance to a choice between two or three items instead of among five, they will gain in the long run by intelligent guessing. Such uncontrollable variations in the informal instructions and discussions which take place in individual classrooms tend to confound the carefully laid plans of examiners and test authors. Correction formulas will discourage such practices and, to some extent, will compensate for the results of such practices, although, of course, it is too much to hope that they will entirely offset them.

Directions to the examiner and to the subjects

As Feder (*25*) has pointed out, the problem of optimum directions needs objective investigation. He states that "builders of standard tests

should recognize the importance of adequate, but not cumbersome, directions, and of determining by experimental procedures the best directions before marketing their products." In the case of a test printed over the years, it is not too much to expect that the directions in successive editions will improve. A professional national jury of experts to criticize each test, before publication, in this respect, as well as in other ways, would be a very useful innovation.

A study by Weidemann and Newens (*69*) indicated that the nature of the directions may have considerable effect upon test scores. They tried out five sets of directions for giving tests involving true-false and indeterminate statements and found significant differences in the resulting scores.

Criteria for preparation of directions.—The following criteria are among those to which special attention should be given when a test maker is writing the directions for the administration of his test:

1. *Assume that the examiner and examinees know nothing at all about objective tests.* Although most teachers and pupils have had experience with objective tests, there are still some individuals who have not used them. Unless the whole procedure of administering and taking the tests is carefully explained, injustice may be done to some individuals, or even to groups of individuals, because of lack of understanding of what is to be done. The test author should take nothing for granted.

2. *In writing the directions, use a clear, succinct style.* Although no steps in the administration of the test should be omitted, long and involved instructions should be avoided. Feder (*25*, p. 30) found that clear-cut, succinct directions were superior to longer, more detailed directions, for the briefer directions tended to avoid the danger of inducing an incorrect mental-set, saved time, and were less fatiguing.

3. *Make the more important directions stand out through the use of different sizes and styles of type.* A common procedure is to use boldface type for everything which is to be read to the examinees and italics for the time limits. Modern test manuals are, in general, much better arranged in this regard than were those of earlier tests. The mechanical arrangement and typography can add to, or deter from, the proper test-taking attitude.

4. *Give the examiner and each proctor full instructions concerning what to do before and after the test is given as well as during its administration.* The manual should make suggestions about advance arrangements, distribution of test materials to examination rooms, instructions to proctors, collection of materials at the close of the examination, and return of blanks to prevent their falling into unauthorized hands.

5. *Check on all possible misunderstandings and inconsistencies by having*

several examiners try the directions out experimentally and report their procedure in detail together with suggestions for improvement. Occasionally, inconsistencies in the interpretation by examiners will occur in connection with instruction manuals carefully prepared under expert supervision. For example, in the older forms of the American Council on Education Psychological Examination, which consisted of five parts, short instructions were printed in the manual and were to be read by the examiner on completion of each part. Time was intended to be taken out for the reading of these directions, but this fact was not specifically stated in connection with the directions. It was found on the basis of replies of schools in a survey made by the Educational Records Bureau that in about one-third of the schools time for the reading of these short directions was not taken out in either some or all groups taking the test. The total time involved was not sufficient to affect the scores of the pupils very significantly, but the inconsistency does lend force to the point that lack of uniformity in procedure is likely to occur in the administration of even a very well-known test unless every aspect of the directions is clarified for all examinees. The work-limit test is less susceptible to such errors than the time-limit test.

As a result of their observation of pupils taking tests, examiners or proctors can sometimes offer very helpful suggestions concerning details which a test maker could not foresee when he wrote the directions. For instance, in connection with the administration of the rate-comprehension part of the Iowa Silent Reading Test, this direction appears in the manual of instructions: "At the end of *one (1) minute* say: 'Stop!' Put a circle around the word you read last, and then continue to read until time is called. You will have two more minutes in which to read as much of this story as you can. Remember, you are to answer questions about it later." One of the users of the Iowa test made the following suggestion concerning this direction: "The student is told to put a circle around the word he read last and to continue reading. Some will do this, while others will listen through the sentences the supervisor has still to read to them, thus destroying uniformity. If 'continue to read' were saved to the last, and if we were instructed to stop the watch during the reading of the directions, the difficulty could be reduced." A longer time limit than two minutes also would reduce the relative error involved.

6. *Keep the directions for the different forms of a test, or for the various booklets in a set of tests designed to be used in the same program, as nearly uniform as possible.* Greater accuracy in the administration of tests can be expected if the examiners and proctors can follow one general procedure for all the tests in a series than would be true if they were

required to use a different arrangement of directions and timing with each separate test. A good illustration of the use of one set of directions to cover a variety of tests in a series will be found in the *Directions for Administering the Cooperative Tests,* one page of which is reproduced in Figure 8. Where the answers are to be recorded in the test booklets, and not on separate answer sheets, the directions for administration of any of the Cooperative tests may be found on this one page.

7. *Where judged necessary or helpful, give practice tests before each regular test.* Such practice tests help to bring about understanding of the test situation and often tend to relieve tension under the actual test administration. Although the practice tests usually will not be scored, if they are easy and are scored, a low score might be used as an index to throw out a test performance as "not sufficiently understood to be allowed to stand as the performance of this individual."

Administration of Tests from the Viewpoint of the Examiner

A well-prepared examiner should be aware of the multiplicity of factors entering into test performance, the kind of environmental conditions, and the special problems created by the current tendency to administer tests with separate answer sheets. He should also have at hand a list of procedures to be followed in preparing for the test, during the test, and after the test has been given.

Factors influencing test performance

A great many variables influence the test performance of an individual pupil. Some of these variables can be controlled during the administration of the test, while in the usual test situations others are wholly beyond the examiner's control.

Factors beyond the control of the examiner.—Among the variables which the examiner cannot control, or cannot easily control, are chronological age; brightness; number of years of schooling; rate of growth; school, departmental, and course objectives; content of courses; teaching methods; teaching ability; practice effect; and tendencies of different teachers and schools to subvert avowed objectives in favor of test objectives.

In establishing the norms for a test, it probably is good practice to discard some test subjects as not being "representative subjects." Toops (59) has proposed an ideal of "replicability of a culture after a decade" as a criterion for eliminating some classes of examinees, or weighting others. Concretely, in 1944, norms of college entrants were decidedly "off-color." The college boys in particular were the boys under 18, not yet subject to the draft; therefore, presumably much brighter than in a more normal

DIRECTIONS FOR ADMINISTERING WHEN ANSWERS ARE RECORDED IN BOOKLETS

Standard Procedure for Administering Tests Not Divided into Parts

1 When all are seated, the examiner should say:

"We shall now pass out the test booklets. Do not open them now. As soon as you get the booklet, fill in your name and the other items of information called for on the cover page. Print your name. When you have finished filling in the blanks, read carefully the directions on the cover page; then wait for further directions. Do not open the booklet until I tell you to do so."

2. Allow sufficient time for filling in the spaces on the cover page and reading the directions. When each student has done this, the examiner may orally emphasize any points that need emphasis, and say:

"Are there any questions? No questions may be asked after the examination begins."

3. Answer all legitimate questions, and then say:

"When I say 'Begin,' turn to the first page of questions, read the directions at the top of the page, and start work. Work as fast as you can without making mistakes. Ask no questions. Read the directions again if you do not understand. You are not expected to answer all the questions in the time limit. Begin."

4. Note the exact time when you say "Begin" and *write it down.* Allow exactly the number of minutes specified for the test, counting from the moment you say "Begin." Do not allow extra time for reading the specific directions *inside* the booklet. At the end of the allotted time, say:

"Stop! Even if you have not finished, close your booklets. See that you have clearly printed your name and that you have given all the other information asked for."

5. Have the booklets collected at once. In doing so, make sure that all the information necessary for identification and classification has been entered. Supply any necessary missing items of information.

Standard Procedure for Administering Tests Having Two or More Parts

Including the Cooperative English Tests, Forms Q Through T

1. When all are seated, the examiner should say:

"We shall now pass out the test booklets. Do not open them now. As soon as you get the booklet, fill in your name and the other items of information called for on the cover page. Print your name. When you have finished filling in the blanks, read carefully the directions on the cover page; then wait for further directions. Do not open the booklet until I tell you to do so."

2. Allow sufficient time for filling in the spaces on the cover page and reading the directions. When each student has done this, the examiner may orally emphasize any points that need emphasis, and say:

"Are there any questions? No questions may be asked after the examination begins."

3. Answer all legitimate questions, and then say:

"When I say 'Begin,' turn the page to Part I, read the directions carefully, and start work. Work as fast as you can without making mistakes. Ask no questions. Read the directions again if you do not understand. You are not expected to answer all the questions in any part in the time limit, but if you should finish before time is called, go on to the next part. If you finish the last part before time is called, you may go back and work on any earlier part. Begin."

4. Note the exact time when you say "Begin" and *write it down.* Allow exactly the number of minutes specified for the part of the test which you are administering, counting from the moment you say "Begin." Do not allow extra time for reading the specific directions at the beginning of the part.

At the end of the allotted time for Part I, say:

"Stop! Even if you have not finished Part I, begin Part II. Read the directions for Part II carefully. If you finish Part II before the time is up, you may go back and work on Part I again, or you may go on to the next part."

5. The examiner should see that all students begin Part II promptly. Allow exactly the specified number of minutes, then say (if there is a Part III):

"Stop! Even if you have not finished Part II, begin Part III. Read the directions for Part III carefully."

6. Thus each part of the test is administered until all parts have been given. Then say:

"Stop! Even if you have not finished, close your booklets. See that you have clearly printed your name and that you have given all the other information asked for."

7. Have the booklets collected at once. Make sure that all the information necessary for identification and classification has been entered. Supply any necessary missing items of information.

FIG. 8.—Directions for administering the Cooperative Tests. Page 2 of an 8-page folder issued by the Cooperative Test Division of Educational Testing Service.

year, say 1940. Norms in 1944 probably would have been more representative if they had been based on women only. Research in this realm is a problem for the future.

Three other factors which may cause considerable variation in the performance of an individual pupil from time to time are physical fitness, emotional state or feeling tone, and motor performance. There seems to be no clear-cut evidence in regard to just how much relationship there is between physical well-being and test performance. It is possible that one's ability to answer test questions is not changed greatly by a headache or a cold, but interest in the test and willingness to put forth effort are frequently affected, and these in turn influence test scores.

Instances are on record where test performance has been changed rather drastically by the emotional status of an individual pupil. Often the examiner is not and cannot be expected to be aware of these subtle causes of unreliability in the results for certain individuals, but in cases where pupils deviate markedly from other data concerning them, the possibility of emotional upsets at the time the tests were taken should be investigated as a possible contributing cause. Such cases should be retested.

As one would expect, there is evidence that motor performance is a factor in scores on a timed test designed to measure abilities other than motor skills. Prichard (*45*) found that in a typical rate test a significant change in writing speed was accompanied by a significant change in score. Motor performance can probably be somewhat, but not wholly, controlled by means of the directions and conditions of the test. Other things equal, the smaller the reaction (check marks) and the less the writing, the better.

Factors which can be controlled by the examiner.—In addition to the variables mentioned, the testing conditions themselves involve a large number of variables, and these are, to a considerable degree, under the control of the examiner. It is not known how much influence lighting, time of day, size of group, and so forth, have upon test scores, but these may be important, at least for certain individuals.

The influence of motivation on test performance was mentioned in an earlier section of this chapter. There is unquestionably a difference among individuals in the attitude and effort they exhibit in a testing situation. There also seems to be a difference in the motivation shown by the same individual at different times. This unevenness of motivation, if it exists in marked degree, can significantly reduce the validity of test results. Without doubt, the difference in motivation from group to group and from time to time for the same individual is due partly to the fact that in a school-wide testing program where a large number of teachers are

used as examiners and proctors, some teachers take an interest in the assignment and do it well, whereas others are bored by and indifferent to the entire testing program, and by their example they destroy the morale of the pupils to whom they give the tests. An interested, businesslike, efficient examiner who is punctual and firm without being too severe can do much to bring about optimum motivation among his examinees.

A question related to motivation concerns whether or not pupils should be warned of a coming test. It seems to be the practice in most schools to inform pupils in advance that they are to take a test, and there is some evidence that this practice is desirable. Tyler and Chalmers (*67*) found that the average scores of junior high school pupils were increased slightly (approximately 1–2 percent) by giving specific warning of the date of the examination two days before it was to be administered.

The seating arrangement in the testing room should be carefully planned both to make the conditions as comfortable as possible for the pupils and to reduce opportunities for copying. Where the size of the room permits such an arrangement, the use of alternate seats, and even alternate rows, is preferable. Individuals otherwise honest are likely to resort to cheating or collusion if a feeling of insecurity over the content of the test is coupled with strong motivation. Fenton (*26*) found, for example, that when given an opportunity to cheat in three experimental situations, 63 percent of a college class of girls did cheat in one or more situations. Bird (*5*) and Dickenson (*19, 20*) have suggested procedures to be employed in studying the number of identical errors in the papers of pupils suspected of cheating. It may be necessary to resort occasionally to such procedures after tests have been given, but obviously the preferable arrangement is to eliminate opportunities for cheating by careful control of the test conditions. Having the seats numbered and requiring the examinees to record their seat numbers may be a deterrent to cheating.

Environmental conditions of the test

In planning the physical conditions of a testing program, the head examiner in each school, of course, has to work within the framework of a particular environment. In some school buildings it is possible to set up almost ideal testing conditions, whereas in others the conditions will be imperfect and will vary from group to group even with the best of planning. It is advisable to give special attention to obtaining the best possible space, lack of crowding in the seating arrangement, adequate light, comfortable temperature and ventilation, and freedom from distractions.

In the average school, the rooms available for testing will probably

include the auditorium, the library-study hall, the school cafeteria, the science laboratory, and classrooms constructed and equipped to handle classes of twenty-five to sixty pupils. One procedure for handling the testing program is to follow the usual school schedule, having the tests administered in the classrooms by the regular teachers. This plan has the advantage of confining the testing to comparatively small groups where it is relatively easy for the examiner to make the directions clear to all pupils and to keep everyone under his personal observation. It is also probable that administration of tests in the regular school routine by teachers whom they know is less disturbing to nervous pupils than the carrying-out of a special testing schedule.

On the other hand, the use of the regular schedule has several disadvantages. In the first place, since the personality, attitude, and procedures of the different examiners constitute a variable in test administration, the larger the number of teachers serving as examiners, the greater the opportunity for the variable to be a cause of differences in test scores. In the second place, some good teachers are constitutionally poor examiners, and one can be sure in advance that they will not do a good job, but they cannot be readily eliminated if the regular schedule is followed. Thirdly, overcrowding and opportunities for copying are likely to occur in the testing of large classes even when pupils are required to move their chairs as far apart as possible.

A fourth disadvantage of following the customary schedule is that different class groups taking such a test as English are likely to be meeting throughout the day. The pupils in classes tested near the end of the day have an opportunity to obtain a certain amount of information and help in connection with the test from pupils who have had it at an earlier hour. Moreover, some teachers who are testing their own classes may be so eager for them to do well that they will yield to temptation to offer a few indirect suggestions which will help the pupils obtain higher scores.

Everything considered, the planning of a special schedule for a testing program usually is advisable, although in the primary grades, testing with regular classes may be preferable. Under a special schedule, the services of the teachers who are potentially the best examiners can be utilized to the fullest extent, a given test may be scheduled at the same hour for all groups, and the best of the school's equipment can be utilized for the testing. A school "holiday" for testing is probably more justifiable than most of the "days" for which holidays are called.

Research is needed on the optimum size of groups to be tested. In the absence of objective evidence, it is not possible to say just how large such groups should be, and, in any event, considerable leeway would be

necessary, for the most desirable group size no doubt depends partly upon the age of the pupils, the ability and experience of examiners in handling large groups, the acoustical properties of the testing room, and the availability of staff members to serve as proctors. If other things are equal, it usually is better to plan for a few rather large groups than for many small ones. When everything is well planned and the room is thoroughly proctored, groups of at least 300 can be tested at one time very successfully by an experienced examiner.

As a rule, the best single room for testing in a school is the library–study hall. This room is likely to be equipped with good desk space and to be well lighted. Some school cafeterias, with their arrangement of tables and chairs, offer very adequate writing space for testing. It is well to keep in mind, however, that often the acoustics in such rooms are not the best. School auditoriums usually are not very satisfactory for test administration because of lack of sufficient desk space. However, those in which the seats are equipped with desk arms which may be raised into place for writing are sometimes suitable for the administration of tests contained in small booklets where all the writing is done in the booklets themselves.

Classrooms equipped with desks, particularly the larger rooms, may also be utilized in the testing program. Rooms with desk-arm chairs should be eliminated from the schedule if possible, particularly where the tests have separate answer sheets.

Reference may be made once more to a point mentioned in an earlier section—that it is desirable for test makers to try to standardize the conditions under which their tests are administered. Little has been done in this direction thus far, and it is true, of course, that as long as many school buildings are inadequate for testing purposes general progress in this direction will be slow. Nevertheless, examiners can themselves do much to improve the reliability of test results by standardizing as far as possible the local environmental conditions under which tests in their own schools are administered.

Administration of tests for machine scoring

Special attention should be given to the equipment of rooms in which tests with separate answer sheets are to be administered, particularly those which are to be machine-scored. Since each examinee must handle an answer sheet as well as a booklet, more desk space is desirable. However, in some large institutions where conditions make it impossible to use large desks in the administration of tests, fairly satisfactory results have been obtained through administration of the tests to all students in a large auditorium under conditions where each student holds a board on his

lap. This procedure tends to bring about comparability within the group even though the results may be somewhat too low as compared with the norms. The surface of the desk, or other writing space, should be smooth and hard in order that the subject may make heavy, black marks without punching through the answer sheet with his pencil and without too much embossing.

When tests are administered with separate answer sheets, occasionally an individual will lose his place and mark a series of answers one place too high or too low on the sheet. This mechanical error, if undiscovered, may lower his score significantly. To forestall this type of error, Taylor (*55*) has recommended a simple device which consists in supplying each examinee with a blank sheet of paper, 8½ by 11 inches, to be used both as scratch paper and as a means of marking his place. He suggests further that one side should be printed with examination instructions so that the examinee will use only one side for notes and thus probably will not transfer graphite from the guide sheet to the answer sheet.

There is evidence that the type of desk influences to some extent the scores on tests given with separate answer sheets. Traxler and Hilkert (*66*) compared the mean scores made on the machine-scoring form of the American Council Psychological Examination by seven pairs of groups of secondary school pupils. Five pairs of groups were selected at random, and two pairs were matched on the basis of Otis IQ. One group of each pair took the test at desks, and the other group took it in chairs with desk arms. All the differences in mean scores were in favor of the desk group, although only one was as much as four times its probable error. Kelley (*34*) applied additional statistical techniques to these data and showed that the difference was clearly significant at the ninth-grade level and that probably the desk group continued to have an advantage in the upper years of high school. These results suggest that if test administrators wish to make sure that pupils have the advantage of optimum conditions when taking tests with separate answer sheets, they should give these tests in rooms having desks instead of chairs with desk arms.

Perhaps the greatest single problem in machine-scoring tests is that a large proportion of the papers often have to be re-marked before they will score properly in the machine. Notwithstanding the reading of special instructions and the use of a short practice test to illustrate the method, many pupils do not understand the necessity of making heavy, black marks on their answer sheets with the special pencil and the need for going over each mark more than once. In the case of some groups of papers received for scoring at the Educational Records Bureau, the staff found it necessary

to re-mark almost every paper before beginning to score them. This extra work tends to create a bottleneck in a testing program with the net result that, despite the potential rapidity of machine scoring, reports of the results may be returned to the teachers little, if any, faster than would be the case if hand scoring were used.

More important than the speed of scoring is the possible influence of wide variations in the excellence of the marking of machine-scorable answer sheets on the validity of the test data for instructional and guidance purposes. There is definite experimental evidence that the mean score of a group of answer sheets marked lightly or marked in a slovenly manner, with many stray dots on the papers, differs significantly from the mean score on a group of well-marked answer sheets (*61*). On an occasional extremely poorly marked answer sheet, the score obtained by machine may be at least 25 percent lower than it should be.

Under these conditions the scoring department is faced with a dilemma. If the answer sheets are scored without re-marking, the pupils with poorly marked sheets will be penalized. This type of penalty may be an effective disciplinary technique, but is an unsound testing technique. On the other hand, if the sheets are scanned and marked so that they will score properly, the conscientious pupils who follow instructions may indirectly be penalized on time-limit tests, for it takes longer to make a heavy, black mark and to go over it several times than it does to make a single light mark.

The problems of machine scoring are not those of scoring alone; they are problems of administration as well. Examiners must cooperate fully with scoring departments in order to make machine scoring accurate and efficient.

One procedure for improving the marking of answer sheets is to provide each room examiner with a sample well-marked paper for exhibition. It is not sufficient merely to duplicate a marked answer sheet by the multilith or photo-offset process. A duplicated answer sheet loses some of the essential qualities of a sheet that will score correctly when inserted in the machine. Each sample answer sheet should be made up by hand with marks that are heavy, black, and *glossy*. The room examiner should pass the sample sheet around before the test is begun and should caution the pupils that if they do not mark their own answer sheets equally well, they will run the risk of receiving scores significantly lower than their true ones.

Proctors should be instructed to move about the room during the examination and to caution individually any pupils they observe who are not marking their answer sheets heavily enough or who are carelessly leaving stray dots and marks on the papers. It is only through constant vigilance

on the part of examiners and proctors that the influence of this variable on the results of machine-scored tests can be eliminated. The number of stray marks on the papers may be reduced if each examiner will suggest to the pupils that they rest their pencils on the question numbers as they work through the test.

Procedures to be followed by the examiner[1]

One member of the staff of a school should have full responsibility for the administration of the testing program. He should carefully plan and carry out all the details of the entire program each year. The following rules for administration of a testing program are based on experience and have been applied successfully.

Preparatory. 1. Select the tests carefully, preferably in cooperation with a faculty committee. If the school is located in an area where there is a state testing program, consider carefully the tests recommended for that program, since they are usually selected by experts in measurement and guidance. Take into account the tests recommended by testing organizations of national scope, particularly when the recommended tests are chosen by committees of administrators and teachers representing different types of schools.

2. Order the tests well in advance of the date on which they are to be used. Allow ample time to get all materials in readiness before the date on which the tests are to start. Check quantity of tests immediately upon receipt, and if more are needed reorder at once.

3. Plan *in detail* for the administration of the tests. Choose examiners and proctors with great care. If possible, use examiners who have had previous experience giving the objective type of test. If inexperienced examiners must be used, they should be carefully rehearsed beforehand. Remember that some very intelligent people are temperamentally unsuited to the exacting routine of administering a test. You may use such persons as proctors for tests being given to larger groups, but they should not be placed in charge of the administration of a test.

4. Duplicate an examination schedule, and see that every person concerned receives a copy. The schedule should give the time and place of each test, indicate just where each class which is to take the test is to go, where the pupils who are not taking the test should be during that time, what material the pupils will need when taking the test, and the name of the faculty member in charge of each examination.

5. Avoid overemphasis on the tests. Urge the teachers to have the pupils take them "in stride."

6. Give pupils who have never taken objective tests an opportunity to examine obsolete editions of tests of the same kind. Better still, have them take a short practice test of the objective type.

[1] Part of this section is quoted from Traxler (*65*, pp. 157-59).

7. Do not distribute the tests to the examiners before the day of the examination. Have packages containing the requisite number of test booklets and all accompanying materials made up and ready for the examiners when the day for the tests arrives. Do this sufficiently in advance that any missing items can be supplied.

8. Provide each examiner with a manual and a sample copy of the test several days before the examination and urge him to study the manual and to practice by taking the test himself. *Most errors in the administration of tests are caused by failure of the examiners to prepare sufficiently beforehand.*

9. Provide each examiner and proctor with a written set of instructions outlining his duties at all stages of the examination.

During the test. 1. Make arrangements so that there will be no interruptions or distractions during the testing period. Persons should not come into, or go out of, the room unless absolutely necessary. This is doubly important with timed tests.

2. Seat the pupils in alternate chairs if possible.

3. See that each proctor understands what is expected of him before, during, and at the end of the examination. The examiner should circulate among his proctors and keep them alert to their duties.

4. Make announcements slowly and clearly in a voice that is loud enough to be heard throughout the room. Assume a businesslike and efficient attitude that will command attention, but do not be unnecessarily severe. Remember, some pupils become nervous when faced with an examination.

5. Have proctors supply all pupils with booklets and pencils and with answer sheets, if the tests are to be administered with separate answer sheets. Announce that the pupils are not to write in the booklets nor to open them until so instructed.

6. Have the blanks on the front of the booklets, or answer sheets, filled out. Be sure to announce the date, specify how names are to be written, and explain other items that may need clarification. Spend sufficient time on this step to see that the information is given correctly by the pupils. Ages and birth dates are especially important on tests of academic aptitude, for these determine what norms are to be employed.

7. Hold faithfully to the exact wording of the printed directions unless there is an excellent reason for introducing a minor variation in them. The preparation of directions for a test is one aspect of test construction and standardization. The wording of the directions has been carefully thought out by the test author. Do not improvise or introduce short cuts. If you do, you may change the test results significantly.

8. Time the examination with extreme care, using an interval timer or a watch which has a second hand and which has been checked for accuracy. It is advisable to have one of the proctors check your timing to be sure that no error occurs. In many tests, accurate timing is the most important single feature of the entire procedure of administration. The proctor should warn the examiner if he gives signs of neglecting his duty, but obviously the examiner must not depend on this warning for his signals.

Timing technique, if ordinary watch is used

1) Synchronize second hand so that it hits 60 when minute hand is *exactly* on a mark.
2) On saying "Go!," look at second hand and *record:* :37
3) Then look at minute hand and *record:* 56:37
4) Then look at hour hand and *record:* 10:56:37
5) Suppose time limit is 7½ minutes:

$$\begin{array}{rr} & 10{:}56{:}37 \\ & 7{:}30 \\ \hline & 10{:}63{:}67 \\ \text{equals} & 10{:}64{:}07 \\ \text{equals} & 11{:}04{:}07 \end{array}$$

6) Glance at watch every minute or two until 11:02 or 11:03. From 11:02 or 11:03, glance at it every 10 or 15 seconds until 11:03:30. Then look at it continuously until 11:04:07, when time is called.

9. Move about the room occasionally to see that all pupils are working on the right part of the examination, but do not stand gazing over a pupil's shoulder until he becomes self-conscious, and do not constantly move nervously from pupil to pupil.

10. Do *not* use the test situation to inculcate good disciplinary habits. The single object of discipline in the test room is to keep everyone working at his maximum all the time, with a minimum of disturbance from all sources, *including* the examiner and proctors. Use gestures, facial expressions, soundless whispers, and so forth, in dealing with examinees during working period. Make it clear that *no* questions will be answered during working periods. If hand is raised, smile and shake head. If anyone speaks aloud or makes semi-audible signs of frustration, smile and put finger to lips; if this persists, frown; if a serious disturbance seems imminent, *remove* disturber from test room quickly and quietly, and make an appointment to clear up the trouble later. Any disciplinary measure which disturbs the group is just as bad as any similar disturbance by an examinee.

11. Stop the examination immediately when the time is up and collect the booklets.

After the test has been given. 1. As soon as a certain test has been given, have all proctors turn in their booklets promptly. Alphabetize and check the papers against the class list.

2. Except in cases of protracted illness, *see that all examinees make up the examination.* This is a bothersome step, but one that is unavoidable, for complete data are essential if the results are to be used successfully in either teaching or guidance.

3. See that the tests are scored promptly. Report the results to the faculty in a form that they can use and provide them with an explanation of the results.

4. Have the scores of each pupil entered on an individual cumulative record card and make this card available to both counselors and classroom

teachers. The card may also be shown to parents if the data are carefully explained in conference. Mature pupils, especially those from the junior year of high school upward, may likewise be shown the results of their tests during interviews.

Scoring of Objective Tests

Among the factors to be taken into account in the scoring of objective tests are the scoring formula to be employed, the weighting of items and parts of the test, the kinds of provision for responses, and the types of keys to be used. Other important factors are the question of whether the responses are recorded in the test booklets or on separate answer sheets, the question of whether the tests are to be scored locally in a comparatively small-scale scoring organization or centrally in a large-scale program, the question of whether the answer sheets are to be hand-scored or machine-scored, the organization of the scoring unit, and the use of special machine equipment. The responsibility for some of these matters rests with the test maker, while for others it is centered in the scoring department. Decisions of test authors in regard to procedures are subject to change by the supervisors of the scoring, if better procedures can be devised provided that they do not affect the scores obtained.

Correction for Guessing

One of the first decisions which must be made about the scoring is concerned with whether the score is to be the number right or whether a formula to correct for guessing and chance factors is to be employed. The generalized formula for correcting for guessing is

$$S = R - \frac{KW}{n-K},$$

where S = score,

R = the number of right responses,

W = the number of wrong responses,

n = the number of suggested responses for a single item,

K = the number of responses to be selected or marked for each item.

Scoring by this formula involves the assumption that every wrong response is the result of a guess, that all wrong responses are equally attractive or equally likely to be selected, and that therefore the law of chance applies to the situation. The logic (or the validity of the logic) underlying the correction for guessing formula has already been discussed on pages 347–51, and need not be repeated here.

In general, most wrong responses are probably due to inadequate

knowledge or ability, rather than to misinformation or wrong learning. The frequency with which a certain wrong response is selected will depend upon the degree to which the item writer succeeded in making that response highly plausible in the light of whatever (inadequate) knowledge or ability the examinee does possess. (See page 349.) If the item writer achieved his aim, the wrong responses will always appear *more* plausible than the correct response to the examinee who does not possess the desired knowledge or ability. Furthermore, the item writer will practically never succeed in making all wrong responses equally attractive, and, thus, the formula will often overcorrect, but to different degrees for different tests or different items.

It is obvious that the greater the number of choices per test item, the less important it is to correct for guessing. Some authorities advise the use of correction formula with items having as many as five choices, and in at least one extensive series of tests—the Cooperative tests—the policy of correcting for guessing is consistently applied. There is some justification for this procedure as far as standardized tests are concerned, for years ago studies by Ben D. Wood (*71*), Eleanor Perry Wood (*72*), Ruch and DeGraff (*47*), and others indicated that while the reliability of corrected scores is not significantly different from that of uncorrected scores, the correction tends to increase slightly the validity of the scores.

In the case of teacher-made tests designed for local use and liberally timed, it is doubtful whether the use of a correction formula is worth the trouble, if the main purpose is to determine the relative standing of the pupils for purposes of marking. Unless some of the pupils omit a disproportionately large number of the items, the correlation between the corrected and uncorrected scores will usually be .98 or .99 on a test that is essentially a work-limit test. The teacher may, therefore, be fairly sure that the pupils will be ranked in approximately the same order regardless of whether or not the scores are corrected. Of course, if the teacher is interested in comparing the scores of the pupils with the highest possible score, true-false and multiple-response tests should be corrected for chance; otherwise the achievement of the pupils will appear to be higher than it actually is. However, if the teacher is not interested in comparing obtained scores with perfect scores, the time spent in applying scoring formulas during the scoring process might better be spent instead in scoring a test 10 percent, or 20 percent, longer than the existing one. Both validity and reliability stand a better chance of being improved by this alternative.

On the other hand, as suggested in a preceding section, many faculty members will prefer to see correction formulas used with all objective tests

in order to discourage what they believe to be a widely prevalent student tendency to form habits of loose thinking and wild guessing.

Among the kinds of objective tests, there is one common type for which the use of a correction formula has seldom been advised. This is the matching test, in which the usual procedure is to employ an uneven number of items in the two series and to score by number right. As various writers have pointed out, however, the formula $S = R$ tends to penalize the more cautious examinees just as it does in multiple-choice tests. Chapman (*8*), Zubin (*73*), Shen (*50*), and Chen (*9*) have discussed the problem of scoring matching tests and have suggested different formulas. Those proposed by the last two writers take their origin from the generalized formula for the correction of scores on multiple-choice tests.

It is obvious, of course, that no correction formula can compensate for differences among pupils in luck at guessing, and these differences may be considerable in a comparatively short test. Aside from psychological and therapeutic values, the whole purpose of a correction formula is to reduce the influences of differences in degree of caution in answering questions.

An interesting variation of the correction for guessing formula is that based on the number of omits (O). The formula is

$$S' = R + \frac{O}{n}$$

where n = the number of suggested responses to each item,

S' = the "corrected" score (not the same as obtained by the usual formula).

The number of wrongs is equal to $W = T - R - O$, in which T is the total number of items in the test. Thus, the usual formula reduces to

$$S = R - \frac{W}{n-1} = R - \frac{T-R-O}{n-1}$$

$$= \frac{nR-T+O}{n-1} = \frac{nR}{n-1} - \frac{T}{n-1} + \frac{O}{n-1}$$

hence

$$\frac{n-1}{n} \cdot S + \frac{T}{n} = R + \frac{O}{n} = S'.$$

Since n and T are constants, S' is a linear function of S, and S and S' are perfectly correlated. Accordingly, since the absolute magnitude of the score is of no significance, the two formulas $S' = R + O/n$ and $S = R - W/(n - 1)$ are functionally interchangeable. The advantage of the S' formula is that the number of omits is usually much smaller and easier to count than the number of wrongs, hence scoring is usually considerably

easier if the S' formula is used. The S' score, of course, is not comparable with the S score (total possible number of rights); the S' scores, although perfectly correlated with the S scores, will be higher and less variable.

Procedures Using Different Scoring Formulas

Where the scoring formula is $S = R$, it is necessary to take account of the right responses only. If the responses are recorded in the test booklet, the scorer marks each right response, usually with a horizontal line, but as a rule he does not mark the wrong responses or the omitted items. If the responses are entered on a separate answer sheet, no marking is necessary, for the right responses can be counted through the use of a cut-out scoring key. For accurate results, a re-count by a second scorer is essential. In the scoring of easy tests, where most of the pupils have answered nearly all the questions correctly, it may occasionally be economical of time to count the wrong answers and omits, and to check the number of right responses by subtracting the total of these two from the total number of items instead of re-counting the correct responses.

Where $S = R - W/(n - 1)$, it is necessary to find the number of right and wrong responses, and it is desirable to take account of omitted items as a check. When test booklets are to be scored, the rights may be indicated with short, horizontal lines, the wrongs with crosses (×'s) and the omits with *O*'s. Each type of mark may be placed in a separate column to facilitate counting. The division of the wrong responses by the proper number to obtain the proportion of the wrongs to be subtracted for guessing is a tedious procedure when done mentally, but simple tables in which the number to be subtracted can be looked up may be prepared and printed in the test booklet or on the scoring key.

One advantage (in addition to that mentioned on page 367) of the use of the correction formula $S' = R + O/n$ is that it is feasible with this procedure to have the pupils do the initial step in the scoring. Before the test papers are handed in, the pupils can be asked to count the number of omitted questions and to record the number on a designated place. In the interest of accuracy, the omitted questions in each test should be counted by two pupils, neither of whom is the one who filled out the test. It also is preferable for them not to know whose test they are scoring.

Still another possibility when the $S' = R + O/n$ formula is used is to have the pupil mark each omitted item, letting the last response position on the answer sheet represent an omit. The stencil scoring key is then prepared as usual, except that a hole appears over every nth *omit* response. For example, in a five-choice test, there will be a hole in the stencil key

for every fifth omit response. A very close approximation to the S' score is then simply the total number of marks appearing through all holes in the key, and the scorer need not count R's and O's separately. The score thus obtained will usually correlate so highly with the true S' score that it is interchangeable with it for all practical purposes.

This procedure is adaptable to machine scoring. It should be observed, however, that if it is used with the usual standard answer sheet, the largest number of choices that can be allowed for individual items is four rather than five.

WEIGHTING

Any scoring procedure whatsoever—in fact the very idea of obtaining a score by summing the data relative to the successful or unsuccessful performance of an individual on a number of unrelated or remotely related items—represents the subordination of a rational philosophy to a practical empiricism. A priori, there would seem to be no justification for adding together a record of one's success on questions relating, let us say, to vocabulary, English usage, number series, verbal analogies, and space relationships, or even on the various items within each of these areas. The only reason why this procedure is allowable is that it results in a score that usually has prognostic value. On a rational basis, the thesis that test items are additive may seem absurd, but, in practice, the idea leads to the provision of useful information concerning individuals and groups.

Once the hypothesis that performance on test items and on the larger parts of a test is summable has been accepted, there remains the question of the relative values of these items and parts. Should some parts have more weight than others? The question of weighting has long presented one of the knottiest of scoring problems, and while there is now rather general agreement concerning unit weighting for most types of tests, the question is still much debated in connection with personality and interest tests and similar questionnaires.

Weighting of items

Theoretically, a rather plausible case can be made out for weighting choices and items in all tests except those which involve the simplest skills. The teachers in any field will insist that some questions are much more important than others and that they should count more heavily in a pupil's score. It can be shown that test questions vary greatly in difficulty and also in validity in terms of correlation with an objective criterion. Any of these considerations, or all three, would apparently justify the use of differential weights in the scoring process.

The authors and users of tests, however, have been reluctant to employ differential weights for test items because weighting when done by hand methods greatly complicates and slows up scoring. For most types of tests the results of research fortunately support their reluctance. For example, Douglass and Spencer (*21*) and Potthoff and Barnett (*44*), and other investigators, years ago established the fact that the correlations between weighted and unweighted scores on objective tests tend to be very high. Douglass and Spencer's correlations for the weighted and unweighted scoring of several tests were .98 to .99. Potthoff and Barnett obtained correlations of approximately .97 to .99. The latter writers said that "for practical purposes the difference between weighted and unweighted scores may be considered to be so small that it may be disregarded, and a great deal of labor may be dispensed with by using the unweighted scores in determining the literal grades." Similarly, Stalnaker (*52*) presented correlations of .97 to .99 between weighted and unweighted scores on essay-type examinations, and concluded that "the relationship between weighted and unweighted scores is so high, so nearly perfect, that there is little justification for the use of weights with these examinations."

Wilks (*70*) showed that, when no independent variable or ultimate criterion is available, the ordinary methods of multiple correlation and least squares in combining tests or other variables are not applicable. He described mathematically three methods of combining variables and discussed them with reference to the linear combination of tests.

On the other hand, in certain instances differential weighting has seemed to constitute an improvement over unit weighting of items. Soderquist (*51*), for example, carried on a study in which the examinee himself weighted his responses to a true-false test according to the degree of assurance with which he made his responses under certain penalty-for-error conditions (2, 3, 4, or double penalty). He found an improvement in the reliability of the test over the same test given and scored under the usual conditions and thought that the weighting procedure had an effect similar to lengthening the test. Hevner (*32*), in a somewhat comparable study, likewise obtained evidence favorable to the weighting of true-false answers on the basis of degree of confidence of the examinee in his own responses.

Notwithstanding experimental results such as those obtained by Soderquist and Hevner, unit weighting of items is now used with nearly all tests except certain interest inventories and personality questionnaires where the responses to different items are characteristic in widely varying degrees of the individuals engaged in a given occupation or possessed of a certain personal quality.

Weighting of parts

In most of the older tests, and in some of the newer ones, the question of the weights to be assigned to the parts of a test is either ignored or handled with very simple arithmetical procedures. Often, the raw scores on the different parts of a test are simply added in order to obtain a total score. Occasionally, where a certain part is regarded by the test maker as especially important, double or triple value is assigned to all the questions in that part. Crude procedures of this kind, when intelligently applied, sometimes result in weighting not greatly different from that achieved by more refined techniques. Such procedures are the only feasible ones for use in the scoring of nonstandardized teacher-made tests containing parts, but in connection with the standardization process, more adequate scoring techniques can be established for published tests.

It should be clearly understood that the parts of a test are always weighted in the total score. The only question is in regard to *how* they shall be weighted. One disadvantage of adding raw scores on the parts to obtain the total score is that the actual weights thus assigned to the parts may not be at all what they appear to be. Superficially, the weights appear to depend on the relative number of items in the different parts. It is true that there is usually a positive relationship between the length of a part and its weight in the total raw score, but this is not necessarily the case. The actual weights of the parts depend upon the relative variability of the part scores and the correlations between them. (See pages 168–69 for a detailed discussion of this problem.)

In the case of achievement tests, no criterion being available, the decision concerning the weights to be assigned to the parts must be made subjectively as a result of an appraisal of relative values of the functions measured.

Where a definite statistical procedure for weighting the parts is employed, the most common plan is to attempt to weight the parts equally. This result is readily accomplished by converting all the part scores into derived scores with the same mean and standard deviation; for example, a mean of 50 and a sigma of 10. If it seems desirable to give more weight to a certain part, the standard deviation for that part may be increased. In the illustration just used, a part could be given approximately double weight by setting the standard deviation for that part at 20, although in that event it might be desirable to set the mean for the series of parts at 100 in order to avoid the possibility that some of the very low standard scores might be negative.

Provision for Responses

The kinds of provision made for the responses of pupils to test items are of great importance in scoring. The efficiency of the scoring process is affected markedly by the types of responses used by the test maker and by his ingenuity in arranging the responses for convenient scoring. However, since the various types of responses are treated thoroughly in chapters 7 and 11, the discussion of this topic will not be repeated here.

Types of Scoring Keys

Scoring keys may be classified as those for manual scoring of test booklets, those for manual scoring of answer sheets, and those for machine scoring. The kind of scoring to be used is, of course, determined in part by the type of objective responses employed.

Keys for manual scoring of test booklets

The most common kinds of keys used in the hand scoring of test booklets are the fan or accordion key, the strip key, the cut-out key, and the transparent key.

Fan, or accordion, key.—Probably the most widely used key for the hand scoring of booklets is the fan, or accordion, key. This type of key is illustrated in Figure 9 with a portion of the key for one of the Cooperative tests. It derives its name from the fact that it may be folded along the vertical lines in the manner of a fan or an accordion to prevent loss of individual keys and to insure their successive use in correct rotation. When folded properly, it may be rotated readily as the scorer proceeds through the pages of a test.

Strip key.—The strip key is similar to the fan key except that the sections for the different pages are each separate and are, as a rule, reproduced on cardboard instead of paper. A strip key can readily be made from a fan key by cutting along the ruled lines, and this procedure is, in fact, sometimes desirable in a large-scale and closely organized program where each scorer is assigned to work on just one page of a test. The Kuhlmann-Anderson Intelligence Test is one of the well-known tests with which strip scoring keys have been used extensively.

Cut-out key.—A cut-out key is one in which windows are cut to reveal words or phrases written into blanks in completion items, or underlined responses in the older type of multiple-choice test. Cut-out keys were formerly used in the scoring of the booklets for a considerable number of tests, but, with the increased use of multiple-choice items and numbered responses, the need for this type of key has declined. If employed, the key

COOPERATIVE ENGLISH TEST
B1: EFFECTIVENESS OF EXPRESSION
FORM S

Raw Score	Scaled Score*
78	90
	89
	88
77	87
	86
76	85
	84
	83
75	82
	81
74	80
	79
	78
73	77
	76
72	75
	74
71	73
	72
70	71
69	70
	69
68	68
67	67
	66
66	65
65	64
	63
64	62
63	61
62	60
61	59
60	58
59	57
58	56
56-57	55
55	54
53-54	53
52	52
51	51
49-50	50
47-48	49
45-46	48
42-44	47
40-41	46
38-39	45
36-37	44
34-35	43
32-33	42
29-31	41
27-28	40
25-26	39
23-24	38
21-22	37
19-20	36
17-18	35
15-16	34
14	33
12-13	32
11	31
10	30
9	29
8	28
	27
7	26
6	25
5	24
	23
4	22
3	21
	20
2	19
	18
1	17
	16
0	15

*The 50 point on this scale corresponds to the expected performance of an "average" individual with 4 years of study of the subject tested at the end of the 12th grade.

The vertical line at the right of the Scaled Score table indicates the Standard Error of Measurement at the 50 point, in terms of Scaled Scores, using parallel-forms method.

See Manual for scoring directions.

Page 2
Col 1
Part I
B1: Expression
FORM S

A(1)
(3)
(2)
b(2)

The total raw score is the number of right answers diminished by 1/3 of the wrong answers. Omitted items are NOT to be counted as wrong. MAXIMUM RAW SCORE = 78 See Scaled Score Conversion Table.

Page 8
Col 2
Part III
B1: Expression
FORM S

24(2)
(4)
(1)
(2)
(4)
29(1)

Page 2
Col 2
Part I
B1: Expression
FORM S

C(1)
(4)
(2)
d(1)

Page 8
Col 1
Part III
B1: Expression
FORM S

19(3)
(1)
(3)
(2)
23(3)

Page 3
Col 1
Part I
B1 · Expression
FORM S

E(1)
(2)
(2)
(3)
(1)
g(3)

Page 7
Col 2
Part III
B1: Expression
FORM S

15(3)
(4)
(1)
18(2)

Page 3
Col 2
Part I
B1 · Expression
FORM S

H(1)
(1)
(2)
(4)
(2)
j(2)

FIG. 9.—Example of fan or accordion key used in hand scoring: single sheet mimeographed on both sides. When folded along the vertical lines, it may be rotated readily as the scorer proceeds through the pages of a test.

should be made of celluloid or other stiff transparent substance. Keys for T-F and multiple-choice tests sometimes have the correct answer positions punched out. These also should be made of transparent rather than opaque material.

Transparent key.—There is considerable use of material upon which the right or wrong responses can be printed so that they may be superimposed upon or placed adjacent to the pupil's answers. Scoring with this type of key, if its transparency is high, tends to be comparatively rapid and accurate, especially when a large number of responses is entered on a single page.

The material most frequently used in the transparent key is paper treated to render it transparent, or plastic material. One difficulty with this type of key is that it tends to roll up and to curl unless heavier, less flexible material is pasted around the edges. A recent test battery, the Chicago Tests of Primary Mental Abilities, Single Booklet edition, employs this kind of key.

One advantage of the transparent key over the cut-out or punched key is that it enables the scorer to scan for double-marked answers at the same time he is scoring. On the other hand, unless the transparency of the material is excellent, the pupil's correct answers may be partially obscured, with the result that the scoring is less rapid and accurate than it should be. It is sometimes advantageous to punch the positions of the correct answers on the key and thus to combine the features of the transparent key with those of the punched key. The Educational Records Bureau uses this type of key to good advantage in the scoring of the Secondary Education Board Junior Scholastic Aptitude Test.

Keys for manual scoring of answer sheets

In the manual scoring of separate answer sheets, the work may be done with punched keys that are plain except for vertical ruled lines to mark off the columns, punched keys with guide lines, transparent keys, or keys overprinted on the answer sheets.

Punched key—plain.—For several of the most widely used tests, the standard scoring keys are punched sheets of light cardboard, unruled except for vertical lines marking off the columns. Among the tests employing this type of key are the American Council on Education Psychological Examination, the Iowa Silent Reading Test, and the Stanford Achievement Test. General observation and the comments of scorers indicate that it is advantageous to punch the key on green stock, which contrasts with the color of the answer sheet and is restful to the eyes. The scoring key for one side of the answer sheet in one of the forms of the American

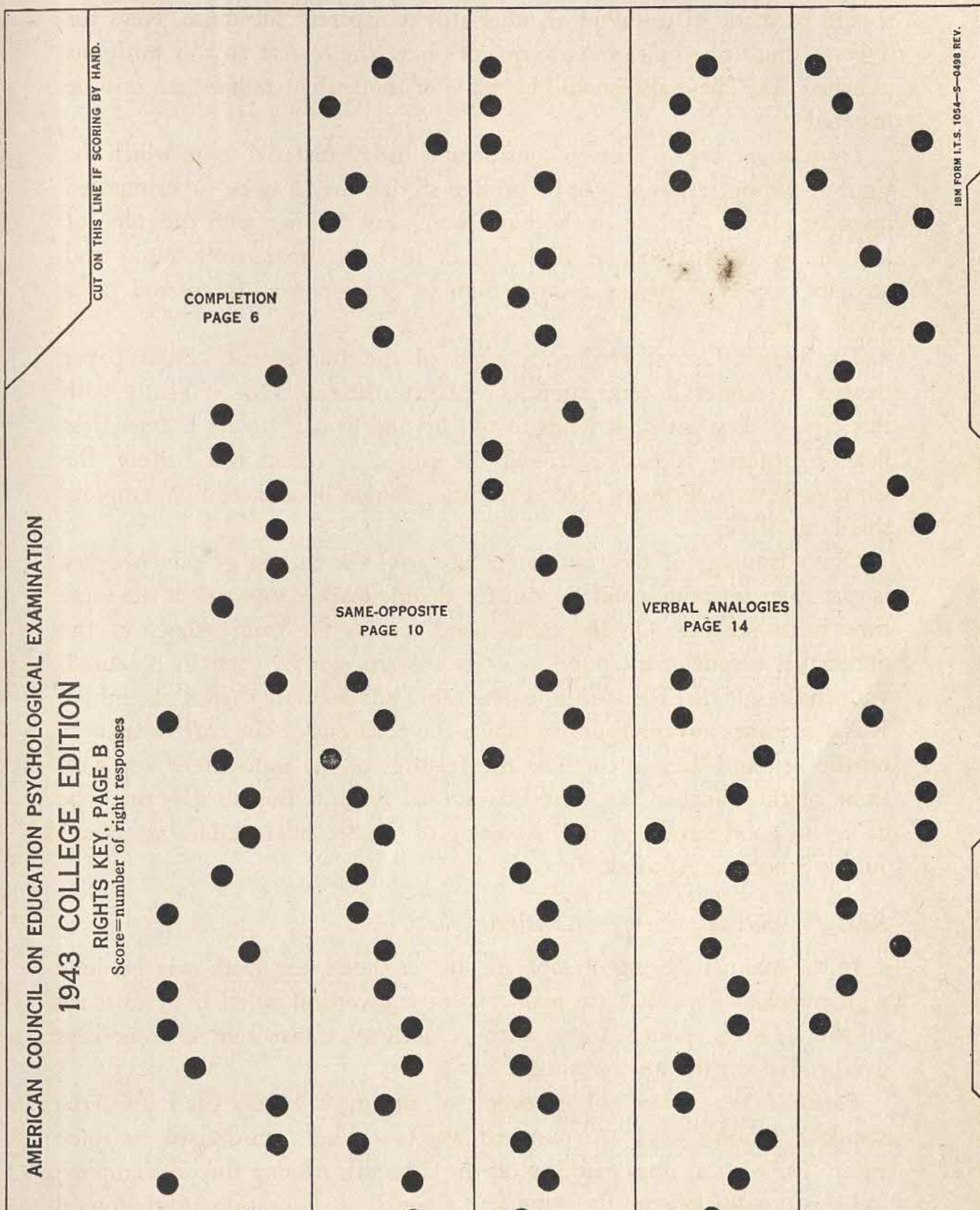


FIG. 10.—Example of plain punched key (without guide lines) for manual scoring. The form is on card stock. The black dots in the reproduction represent holes in the original form.

Council Psychological Examination is shown in Fig. 10. Scoring can be done very rapidly with this type of key, particularly when the score is simply the number right.

Punched key with guide lines.—A slightly different type of punched key has a printed line to guide the eye from one response position to the next one and does not have the vertical lines. Some test experts have indicated a preference for this type of key, and it has been used in some very large-scale testing programs such as those conducted with the Iowa Every-Pupil Tests of Basic Skills and the Army-Navy College Qualifying Tests. Apparently there are no experimental data on the relative efficiency of this type and the preceding one. However, in one of the regional offices for the Army-Navy College Qualifying Tests, an informal experiment with these two kinds of keys was made near the beginning of that program. The criterion was simply the preference of the scorers after they had worked with the two types. In the end, all the scorers in that particular office who tried out the two kinds were unanimous in preference for the one with the guide lines. The value of guide lines no doubt depends partly upon the compactness of the key. Experience indicates that where the items and the responses are printed close together guide lines are not desirable.

If guide lines are used, light lines appear to be preferable to heavy ones. A black line about one thirty-second of an inch wide on a light-green, dull-finish background seems to work best. For this key and also the preceding one, alignment marks may be printed on the answer sheet and corresponding holes punched in the key. A desirable alternative is to cut off the upper left-hand corner of the answer sheets and also of the keys after the manner of IBM cards. A still more efficient alignment procedure is to use a scoring frame, and to have the answer sheets cut and printed so that the copy is always perfectly registered to the left and bottom edges of the sheet. See page 379 for further discussion of this procedure and the scoring routine used with it.

Transparent key.—The use of a transparent key in scoring answer sheets probably is not advisable, for the incorrect responses cause too much visual interference. The only point in its favor is that this type of key might eliminate the scanning step which is always necessary with the two preceding types, but its disadvantages apparently outweigh this one advantage.

Overprinting.—In lieu of a separate key to be superimposed on the answer sheets, the key itself can be printed with a duplicating machine directly upon the answer sheets after they have been filled out by the pupils. The labor-saving advantages of the overprinting of answer sheets have been urged for some years by various test specialists. Toops employed

this device in 1930 on Form 17 of the Ohio State University Psychological Test. The overprinting was done by an electroplate on both sides of the answer envelope, and 296 questions were scored with only two press impressions. In 1935 Stenquist (*53*) showed how the labor of scoring was reduced by this device in connection with the tests of the Baltimore Public Schools. In connection with the Iowa Tests of Educational Development, Lindquist experimented with ingenious answer-sheet arrangements and overprinting patterns to facilitate the speed and accuracy and to reduce the cost of scoring. Encouraging results were obtained, but no kind of overprinting was found to be as satisfactory as the use of a separate scoring key and a scoring frame.

A Multilith is probably the best equipment for accurate registration in the overprinting of answer sheets. In connection with the spring 1946 and 1947 testing programs for independent schools, the Educational Records Bureau overprinted thousands of answer sheets for the Cooperative English Test on the Multilith without a hitch of any kind. (See Figure 11.) It was found that on the complete English test, the scoring time saved by this procedure in comparison with hand scoring of test booklets was 6.4 minutes per test, or approximately 42 percent. It has been found helpful to scorers to overprint in colored ink, preferably green or brown.

Few schools are equipped with Multiliths, while most schools do have Mimeograph machines. There is some evidence that overprinting with a Mimeograph is feasible. Experimentation with the Mimeograph at the Educational Records Bureau indicated that the registration was sufficiently accurate on about 99 percent of the answer sheets (*63*). One practical difficulty is that the IBM answer sheets for certain tests are too wide to be fed through a standard Mimeograph.

One advantage of overprinting is that this procedure to some extent meets the objections of teachers who say that in the usual procedure of scoring separate answer sheets, the right and wrong answers are not marked, and thus the answer sheets do not serve the purpose of review and reteaching in the same way that scored test booklets do. The overprinted answer sheets, when returned to a class, do enable the teacher and the pupils to discover quickly which questions were answered correctly and which ones were missed. This point will be evident on examination of the overprinted answer sheet shown in Figure 11. When machine scoring is used, the answer sheets may be overprinted *after* the scoring process is completed. It has been found that this procedure increases the usefulness of the scored tests for the teachers.

In lieu of overprinting, other rapid devices for showing the correct responses on answer sheets have been suggested by various persons. Bice

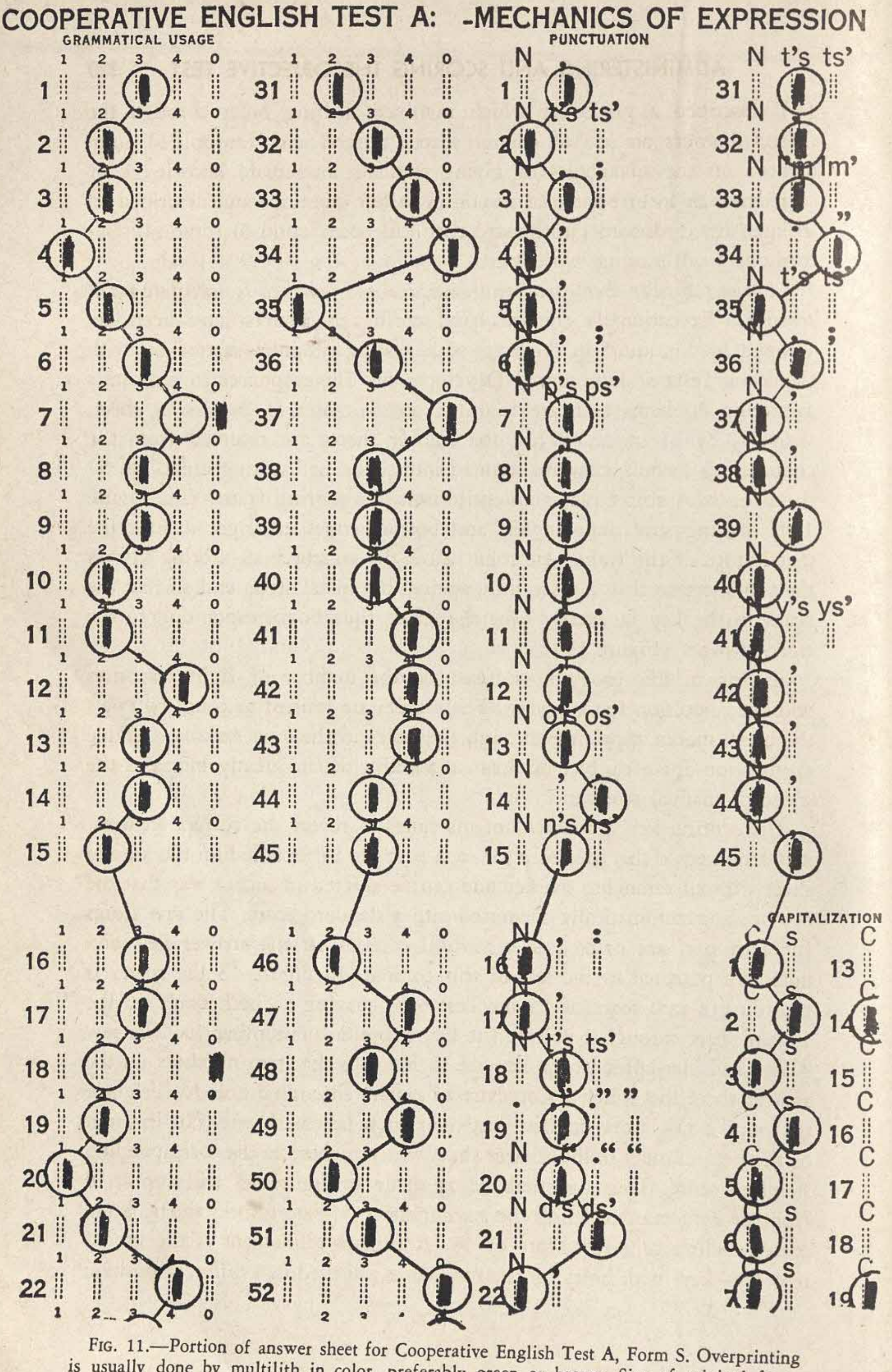


FIG. 11.—Portion of answer sheet for Cooperative English Test A, Form S. Overprinting is usually done by multilith in color, preferably green or brown. Size of original form, 9¼ × 11 inches. Raw scores at left, with name, class, and other pertinent data on student.

(*4*) described a procedure which involved drilling holes through the correct answers on sets of answer sheets. Angell and Troyer (*1*) commented on the advantages of giving students immediate knowledge of correctness or incorrectness of responses to test questions and described an inexpensive cardboard punchboard that had been found to serve satisfactorily as a self-scoring test device.

Compact answer sheet, scoring frame, and key punched to facilitate rapid scoring.—Exceptionally efficient hand-scoring procedures have been developed by Lindquist in the large-scale testing program carried on with the Iowa Tests of Educational Development. The responses to nine tests averaging 80 items each are recorded on two sides of an answer sheet, size 8½ by 11 inches. When the answer sheets are received from the cooperating schools, they are counted into packs, or "scoring units," of 50 sheets each. A scorer places an entire unit in a scoring frame (see Figure 12), making sure that the left and bottom edges are tight against the metal stops of the frame. He then places the punched-out scoring key on the top sheet so that it fits tightly against the metal stops and so that the holes in the key fit exactly over the small squares corresponding to the right answers (Figure 13).

Scoring in this program is based on the number of right responses without correction for guessing. The scorers are trained to count by two's the black marks appearing through the holes in the key, making a single eye-fixation for each pair of dots—a practice which greatly increases the speed of manual scoring.

The scoring key not only contains holes to reveal the correct answers; it is arranged so that the score for each part can be recorded on the answer sheet without removing the key and can be entered in such a way that the raw score is automatically converted into a standard score. The raw scores for each part are printed in a vertical column on the answer key, and holes are punched to the left of this column. In Figure 13 the scorer is recording a raw score of 59 on Test 3 by making a check mark on the answer sheet through a hole in the key opposite the number 59. The raw scores and the holes in the key are so fitted to the item numbers on the answer sheet that when a scorer makes a mark through a hole to designate the pupil's raw score on that particular part, the mark will fall in front of the item number in the answer sheet which represents the corresponding standard score. Thus the mark being made in Figure 13 will appear in front of item number 15 on the answer sheet, 15 being the standard score corresponding to a raw score of 59. A further refinement is the use of rescorers' keys with holes for the raw scores punched in a different position

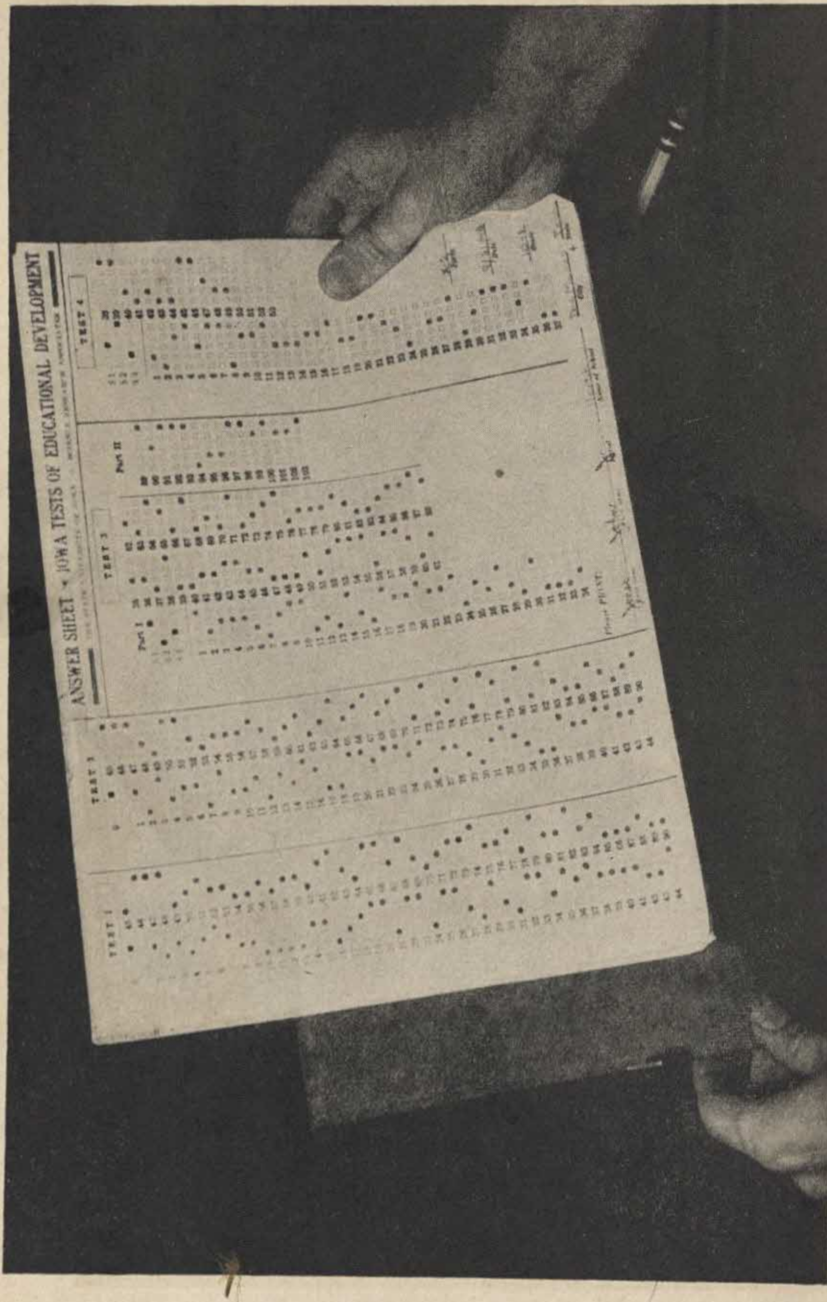


FIG. 12.—Answer sheets are placed in scoring frame

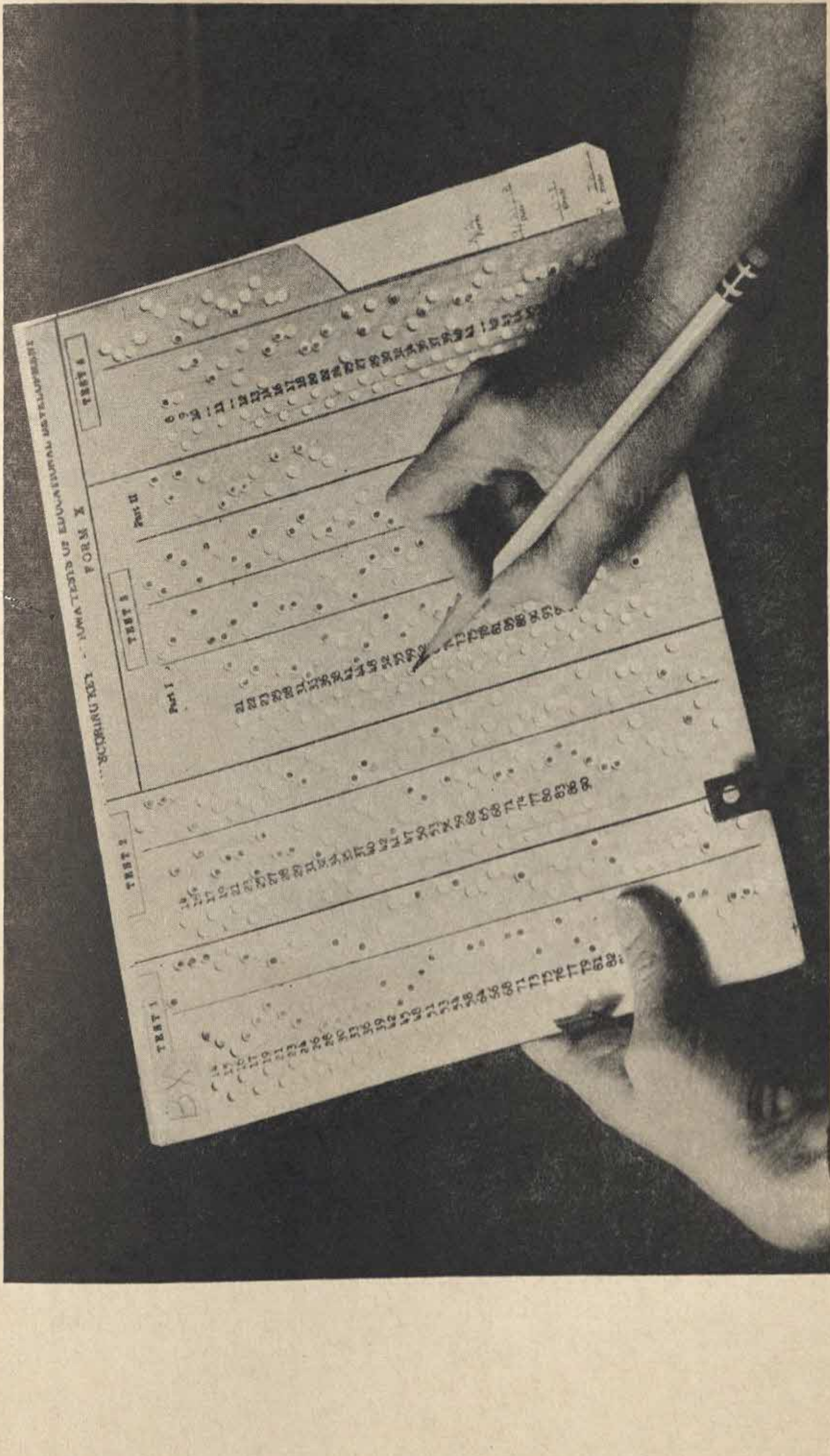


FIG. 13.—Scoring with stencil key

from those on the original scorers' keys, so that the second scorer cannot see the marks made by the first scorer.

As a scorer completes the scoring of a side of an answer sheet, he simply slips the answer sheet out from under the scoring key and lays it face up in a cardboard box to his right. The next answer sheet is exactly in position for scoring, and thus the scorer can work through the pack of 50 sheets without loss of time in adjusting the scoring key with the answer sheets. When one side of all 50 answer sheets in the scoring unit has been scored, the scorer turns the pack over, reinserts it into the scoring frame, takes a new scoring key, and repeats the process.

Some data on speed of scoring under this arrangement are given near the end of this chapter.

Keys for machine scoring of answer sheets

The same punched key (plain) that is used in hand scoring (Figure 10) may also be employed in the scoring machine if the answer sheet has been arranged for machine scoring and if the key is cut properly and is punched with holes of the right size. The function of the key is to divide the contact points into one circuit for the right responses and one for the wrong responses. In addition to this key, an "item-elimination" key is sometimes used in order to block out certain fields. An ingenious use of an item-elimination key is to throw the right responses of a part of a test upon the wrongs circuit so that the rights can be scored as wrongs. This types of use is feasible with tests for which four to six part scores are obtained from one side of an answer sheet and in which the score is the number of right answers uncorrected for guessing. Otherwise, not more than three part scores can be obtained from a single side in one insertion of an answer sheet.

Most machine-scoring keys are now punched on light cardboard stock just as the hand-scoring keys are. These keys, while fairly substantial, do not stand up well under long and continuous wear. Composition material possessing insulating properties, such as bakelite, has been used in a few instances with very satisfactory results and will probably be employed more extensively in the future.

THE USE OF SEPARATE ANSWER SHEETS AND ANSWER PADS

The use of separate answer sheets and answer pads with objective tests is based on the simple but fundamental idea that a subject's choice among several alternative responses may be represented by a mark placed in a numbered position to agree with the response which the examinee con-

siders to be right. Five answer positions of the size used on the IBM answer sheets may be placed in one inch of space, allowing as many as seven pages of test material to be scored on one page. With the compact answer sheets employed with the Iowa Tests of Educational Development, 54 closely printed test pages are scored on two sides of a single $8\frac{1}{2}$ by 11 answer sheet. In view of the simplicity and inexpensiveness of this kind of arrangement, it seems surprising that objective tests were used for nearly a generation before the use of separate answer sheets became at all general. It is true that years ago separate answer forms were provided for certain tests, notably the Ohio State University Psychological Test, the Henmon-Nelson Tests of Mental Ability, and the Nelson-Denny Reading Test, but it was not until about 1936 that answer sheets were made available for a large number of tests. The development of machine scoring gave great impetus to the preparation of answer sheets for standardized tests.

Advantages and limitations

Answer sheets or pads have several advantages over test booklets as a medium for the marking of the responses of the examinees. They are inexpensive, and they make it possible for the same test booklet to be used repeatedly, since the subjects do their writing only on answer sheets. They permit the recording of 150 or more responses on a single sheet of paper. They contribute to the speed and accuracy of hand scoring by eliminating much turning of pages, reducing the amount of eye movement, and eliminating completely the time-consuming step of marking the right and the wrong responses on the examinations. From the standpoint of speed, accuracy, and economy of scoring, the recording of responses on the answer sheets is unquestionably preferable to the writing of responses in test booklets.

On the other hand, the use of separate answer forms presents certain disadvantages. Although the scoring of separate answer forms is more efficient than the scoring of booklets, some scorers say that they find answer-sheet scoring more fatiguing and more conducive to eye strain than booklet scoring. This objection on the part of scorers is not very prevalent when the score is simply the number right, but it is frequently voiced when the scorers must count wrong responses as well as right ones in order to correct for guessing. If many individuals on the scoring staff are thus affected by the answer sheets, the morale of the staff may be somewhat lowered.

In the section on administration of tests, reference was made to another possible disadvantage. It is that the necessity of adjusting the answer sheet

to the test booklet when the test is being taken introduces an extraneous variable, apparently somewhat dependent upon the motor skill and space perception of the individual, as well as on the desk space provided. This variable is probably of little importance where the answer form is closely coordinated with the test booklet, as in the Ohio State University Psychological Test where the back-cover page of the test booklet is an envelop to guide the sheet and to keep answers and record spaces in approximate alignment as successive columns of five answer spaces are brought into use. On tests for which there is not equal spacing of items in booklets and answer sheets and for which no mechanical guides to adjustment are provided, the influence of this factor on the scores of some individuals may be considerable. This is another question on which research is needed.

The experimental evidence in regard to the influence of answer sheets on the reliability and validity of scores is not extensive. Two master's theses on the use of separate answer sheets with achievement tests in the intermediate grades were written at the University of Iowa. In one of these, Herkelmann (*31*), using Forms A and B of the Iowa Elementary Language Tests with 266 pupils in grades four to six, found no significant difference in mean scores with and without answer sheets when the tests were administered without a time limit, but there was a significant difference in the time required. He estimated that the correlation between the scores on the two equivalent forms, when one was administered with and the other without an answer sheet, was considerably lower than the reliability of either form when administered in the usual fashion. However, the reliabilities for the separate forms were computed by the odd-even items method. They would, therefore, be expected to be considerably higher than the correlation between forms, regardless of the effect of the answer sheet. In the other thesis, Loper (*39*) found more time required for the use of the separate answer sheet than the test booklet in grades three and five, but no significant influence on the mean scores when enough time was allowed. There was no evidence in this study that the use of separate answer sheets reduced the reliability of the scores.

The most extensive available study of the effect of separate answer sheets is one reported by Dunlap (*22*), who carried on a series of five experiments in which the use of answer sheets was compared with underlining the correct answer in terms of means, standard deviations, reliability, and validity. The pupils were in grades four and eight. In general, the data indicated that answer sheets were as satisfactory as the underlining method. Dunlap recommended the use of an articulated, serially numbered answer sheet, particularly if the test was short enough to permit the record-

ing of all answers on a single side of the sheet. Careful attention in printing format will reduce errors in use of separate answer sheets. For instance, in Form 18 of the Ohio State University Psychological Test, the different pages of answers were purposefully thrown out of alignment to warn the examinee if he attempted to use the wrong answer page of the record. Science Research Associates has also made use of the articulated answer sheet in a number of recent tests. An answer sheet or answer pad is articulated to a "step-down" booklet (one in which the pages become successively narrower from front to back) and is held firmly in place so that the danger of getting the booklet and answer sheet out of alignment is virtually eliminated.

One disadvantage of separate answer forms that some teachers and counselors stress considerably is that the scored sheets are less useful than scored booklets in diagnosis, teaching, and counseling, because they do not show what questions each individual answered correctly. As already suggested, this criticism can be met to a considerable degree by overprinting or otherwise designating the correct answers on the pupils' answer sheets, or by the use of the graphic item counter or tabulation by hand of the most prevalent type of error.

Range of applicability

Separate answer forms are apparently suitable for use from grade four upward. Dunlap's study (*22*) did not indicate that answer sheets were noticeably less satisfactory in grade four than in grade eight. In fact, Loper (*39*) found that after practice answer sheets caused no difficulty even in grade three, although the third-grade pupils had some difficulty the first time they used the separate answer sheets.

Kinds of answer forms

The kinds of answer forms may be classified as carbon paper, stylus, or pin prick, hand scoring with punched key or overprinting, and machine scoring.

Carbon paper.—Among the answer sheets first used with objective tests were those in which carbon paper was employed. This type of answer sheet is best represented by the Clapp-Young Self-Marking Tests, including the Henmon-Nelson Tests of Mental Ability and the Nelson-Denny Reading Test. The answer sheet is folded and the edges sealed together. The outside pages are used for the writing of the responses. The individual records his response to each item by making a cross (×) in one of five small numbered squares. The answer sheet is spaced to correspond with

the items in the test booklet, and the adjustment is facilitated by means of arrows. The squares representing the position of the correct answers are printed on the hidden inner side of the sheet, and strips of carbon are so placed that they reproduce the mark made by the pupil. The cross falls within the square if the pupil's choice is correct, and outside it if it is the wrong response. Apparently no studies comparing the speed of scoring this type of answer sheet with the others have been made. One would estimate that it can be scored as rapidly as an overprinted answer sheet and perhaps a little faster, for there are no spaces for wrong responses which might interfere with eye movements. The main disadvantages of this kind of answer sheet are the extra expense and the messiness of carbon, which is likely to smudge. It is doubtful whether this kind of folded sheet containing carbon paper can be produced as cheaply as a single sheet designed for scoring manually or mechanically.

Pin-prick.—Partly because the Clapp-Young people had patented the carbon-paper device, other test makers experimented with the idea of having the subject indicate his choice by punching a pin through an answer pad. This technique (patented) has been successfully employed for years in the Ohio State University Psychological Test. It is utilized in recording responses to a widely used interest test, the Kuder Preference Record, as well as in a variety of other tests distributed by Science Research Associates. In each of these tests, the answer pad is closely articulated with the test booklet, and thus a minimum of motor adjustment is required of the examinee. The pin-punch procedure is especially well suited to the Kuder record, since in this way the subject records his responses on all nine scales at once, and the scoring boxes on each sheet are adjusted to the scoring of the "interest" in question.

The Toops scoring pad, which is used with the Ohio State Psychological Test, contains three identical answer sheets, each identified by a common serial number. This arrangement makes it very convenient for the examiner to have a group of students each score three papers immediately after the test has been taken and thus to check and double-check independently the scoring, thus practically eliminating the danger of scoring errors. When the scoring is done by clerks under supervision, only two of the three sheets need to be scored (38). The answer sheet which is to be "pricked" by the pupil and one of the underlying (hidden) scoring sheets from the Toops answer pad are shown in Figure 14. The pin-prick answer pads, like the carbon paper answer sheets, are somewhat more expensive than the single answer sheets. This is the only device which produces multiple identical copies of the test responses for filing in different school offices.

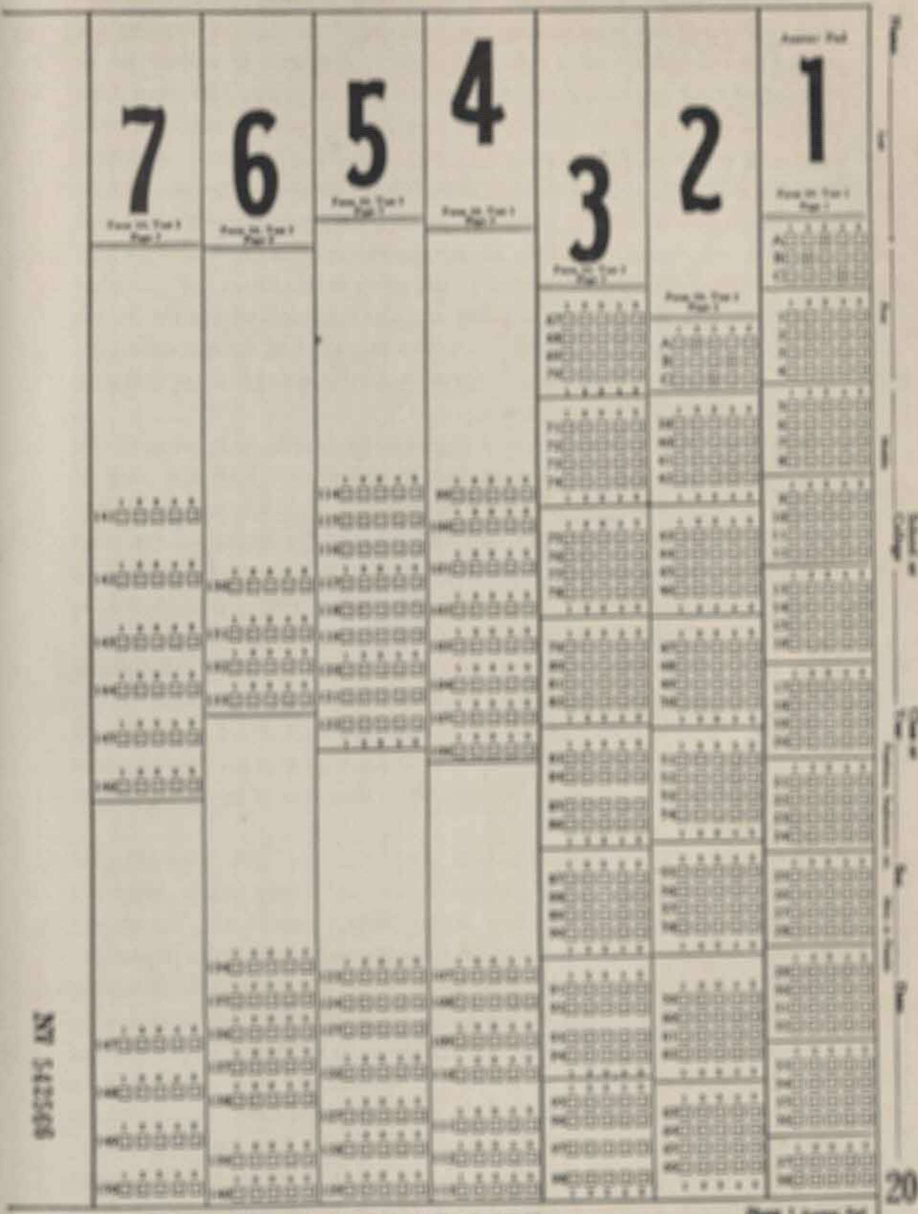


FIG. 14.—Answer sheet of Toops Scoring Pad, Form 20. Answer sheet is pricked with pin by pupil; underlying (hidden) is the scoring sheet, which is shown on the other side of this page. Original form comprises four sheets 8½ × 11¾ inches. Reproduced by permission of H. A. Toops, Ohio State University.

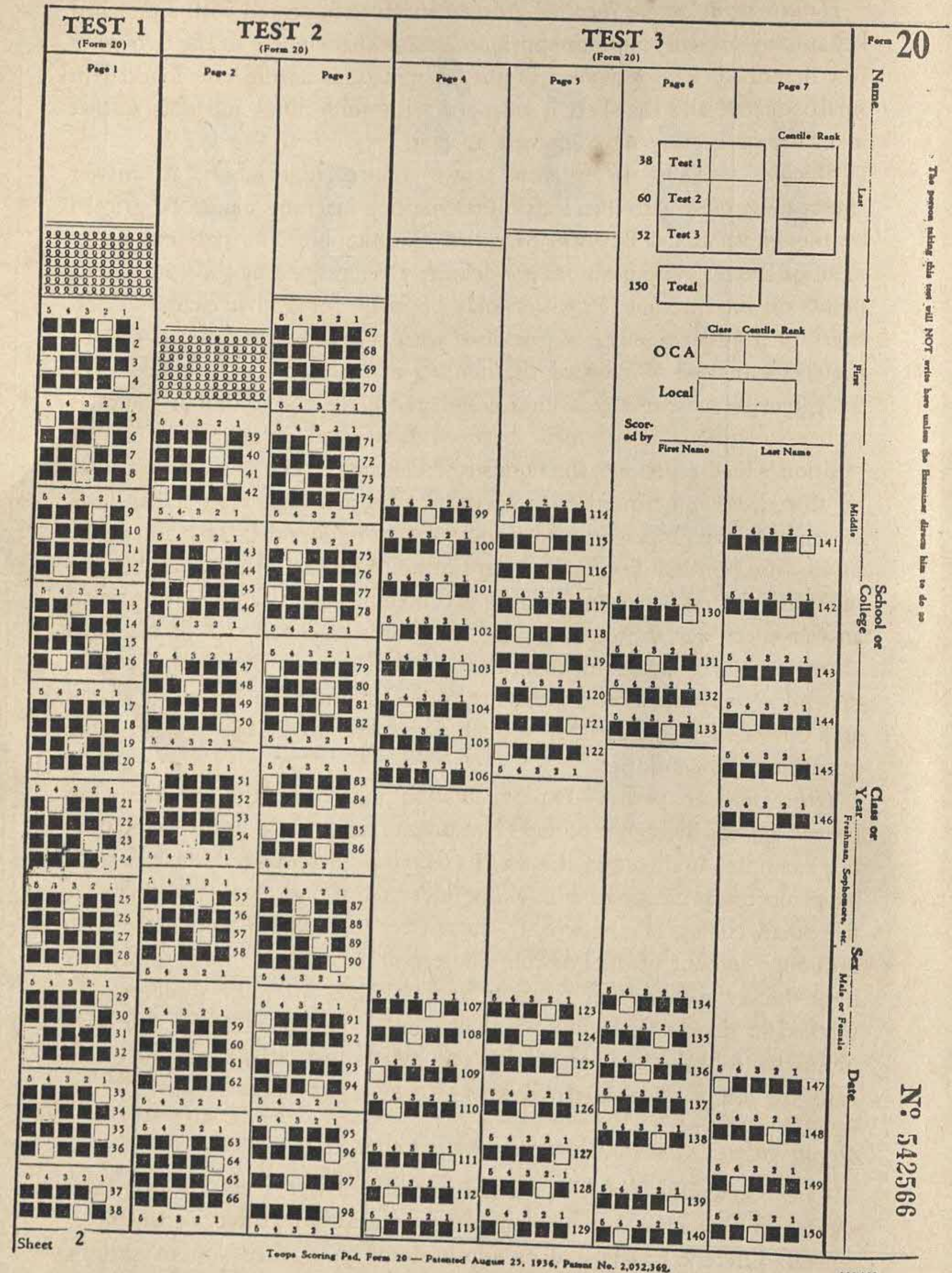


FIG. 14.—Form 20 (reverse side)

Hand-scoring single sheets.—Answer sheets to be scored with a punched key usually present a surface appearance somewhat similar to the two types just described. The responses of the examinee, as a rule, are marked in small squares, and the sheet is arranged with guide lines and dots so that a key can be quickly superimposed upon it.

Machine-scoring.—In order to assure accurate registration, all answer sheets to be used with the electrical test-scoring machine should be printed by the International Business Machines Corporation. The size and placement of the response positions are definitely determined by the 750 contact points on the machine. Provision may be made for 150 five-choice questions, or a larger number of questions with fewer choices, on one side of the sheet. Instead of showing the number of the response he chooses for each item by making a cross in a designated place, the examinee indicates it by penciling a heavy mark between two small printed lines in the position which represents the number of his choice. The answer sheet may be scored either by inserting it in the scoring machine after it has been set up with the proper key and reading the score on a dial, or manually by means of a punched key, or by overprinting. A special pencil *must* be used in taking the test if machine scoring is contemplated. A sample overprinted answer sheet was shown in Figure 11. The usual retail cost of machine-scorable answer sheets in quantities of 25 is two to four cents a copy, depending on whether the sheets are printed on one side or on both sides and on whether standard or special sheets are used. The cost in large quantities is much lower.

Other answer sheets.—Experimentation in an attempt to devise improved answer sheets is reported from time to time in educational literature. For example, Wallen and Rieveschl (68) have described an answer sheet based on the principle of ultraviolet fluorescence. For each item there are five small circles, one of which (the correct choice) is impregnated with a minute amount of an invisible fluorescent substance. Under ultraviolet light the errors stand out as bright spots of light after the students have blacked in the circles to indicate their choices. It is claimed for this technique that it avoids the need for special styli and special pencils and eliminates the use of separate answer keys or cumbersome answer pads.

Organization of answer sheets for test batteries

One noteworthy feature of hand-scorable answer sheets is that these sheets can be arranged to accommodate an entire test battery consisting of several different booklets to be administered in a succession of sittings. This arrangement is very advantageous for speed of scoring, and it elimi-

nates the necessity of collating several separate answer sheets for the same individual.

For an extensive test battery covering the work of a school year, an answer folder has usually been required, but Lindquist has devised an answer sheet for use with the Iowa Every-Pupil Tests of Basic Skills which accommodates 14 tests, including a total of 701 items on the two sides of an $8\frac{1}{2}$ by $12\frac{1}{4}$ inch page. Scoring is done manually through the use of a punched key, in the manner described on page 379.

Use of "homemade" answer sheets

Since separate answer sheets have not as yet been published for a considerable proportion of the standardized tests, a question arises in regard to whether it is feasible and legitimate for a school to make up its own answer sheets for various tests and to duplicate them for use by its pupils. A plan of this kind could be put to use with a minimum of expense. In schools which never have and apparently never will be able to afford a sufficient quantity of regular test materials to supply their pupils, the local preparation of answer sheets would not encroach upon the publisher's welfare and might even be an advantage to the publisher since it would lead to the sale of a small number of tests where none were sold before. For schools whose budgets permit the purchase of regular test materials from the publishers, the ethics of duplicating locally made answer sheets may be questioned even though the legality of the procedure is clear. It would have unfortunate consequences for test users if "homemade" answer sheets were so extensively employed in schools throughout the country that the sale of standardized answer sheets was greatly reduced. Most tests have to pay their own way and publishers would hesitate to issue new tests if it seemed likely that the sale would be too small to pay the cost of publication. Thus, wide adoption of the practice of using locally prepared answer sheets with standardized tests might eventually retard the development of improved objective tests.

A further limitation to the use of locally made answer sheets with time-limit tests designed for the recording of responses in test booklets is that the resulting scores of class groups will tend to be somewhat too low for direct comparability with the publisher's norms, since it takes longer to enter responses on separate answer sheets.

Addends as a potential means of scoring

If the examinee's pattern of response to a number of test items can be indicated by a code number, one can score several items at once by referring the code number to a scoring book.

Example 1

	T	F
Question 1?	0 ☐	1 ☒
Question 2?	0 ☒	2 ☐
Question 3?	0 ☐	4 ☒
Question 4?	0 ☐	8 ☒
Question 5?	0 ☒	16 ☐
		$\boxed{13}$

Code number of the questions marked in this section. Add the numbers before the boxes you need.

Pattern 13, for a given test, has five rights if the stencil agrees with the ×'s of the example above; it has two rights if all the questions are all true.

Example 2

	1	2	3
Question 1?	0 ☐	1 ☒	2 ☐
Question 2?	0 ☐	3 ☐	6 ☒
Question 3?	0 ☒	9 ☐	18 ☐
Question 4?	0 ☐	27 ☒	54 ☐
			34

Code number of the questions marked in this section. Add the numbers before the boxes you need.

With the alternatives themselves numbered with addends, the boxes shown above can be dispensed with and the addends may be employed as a response. Pupils can exchange papers to verify the adding of the addends. Reference to prepared tables, in which all possible patterns have been scored, reveals the score on all the items. There are 32 patterns, so, to be scored for the T-F test above, coded 0–31 inclusive; and 81 for the three-choice test. The addends must be added errorlessly or serious scoring errors may occur.

Toops (57) has proposed addends as a means of rendering mechanical performance tests amenable to IBM scoring. The examinees record the code numbers of their performance, have them checked by a seat-mate, and then dissemble the parts, ready for administration to a new examinee.

Organization and Work of a Machine-Scoring Unit

Machine-scoring units vary in size from those in individual schools and colleges equipped with a single machine to the larger test service organizations which may be supplied with half a dozen or more machines. Since

the annual rental, including the cost of servicing, is about $500, either a fairly large school or a cooperative arrangement among several schools in the same area is required to make the installation of even one machine a paying venture. Where possible, it is desirable for a scoring department to have at least two machines, for this arrangement makes it possible to have the answer sheets scored in one machine and rescored in the other, thus checking on mechanical faults in each machine as well as on operator accuracy.

When a single scoring machine in a school is being operated on the usual part-time basis as needed in connection with the school's regular testing program and also for the scoring of teacher-made tests, an operator and one adult assistant, or two student assistants, should be able to handle all the scoring except during the fall and spring peak loads, when additional part-time assistants may be needed. Thus, schools will not, as a rule, find the expense for personnel to operate the machine prohibitive.

The organization of a machine-scoring unit involving two or more machines calls for a supervisor who is also an expert operator, preferably at least three assistant operators, and not less than three assistants to scan the answer sheets before scoring, hand-check the answer sheets showing discrepancies as a result of machine scoring, make distributions of scores, and so forth.

When a group of answer sheets is received for scoring, the following steps should be taken as a minimum:

1. Classify the answer sheets into the desired age, class, sex, or other groups. Hand-stamp to prevent mistakes [illegible].
2. Alphabetize them, and check against class lists or information cards.
3. Scan each answer sheet for discrepancies or incompleteness in the filling-out of the information blanks, double-marked items, marks too light to be picked up by the machine, and stray dots and other marks. Darken marks and erase as needed.
4. Score the answer sheets in one machine and rescore in the other. Alternate operators between scoring and other kinds of work, rotating shifts of about one hour duration.
5. Hand-score all answer sheets which show machine-scoring discrepancies.
6. Make distributions of the scores in the different groups and compute medians, quartiles, and ranges.
7. Score percentiles on answer sheets, if that service is desired.
8. Have class lists and lists of scores typed if that service is desired.
9. Make graphs with the item counts of the correct responses to the various items in the different groups, if desired.

A very high standard of accuracy, good motor coordination, and quick-

ness of visual perception are qualities needed in a scoring-machine operator. The scoring machine is not difficult to operate, and the training program for an operator is relatively simple. It is advisable to start a new operator on the scoring of tests where the scoring formula is $S = R$. The machine is customarily set so that the needle on the dial registers halfway between two marks, and the operator reads the lower one. When the machine is set properly and is in perfect working order, the registration for tests in which the score is the number right will be almost exact throughout the scale. When a correction is made for guessing, however, the operator has to exercise judgment with regard to the dropping or retention of fractions of a point, depending on just where the needle comes to rest. Experienced operators become very expert at making these judgments, and usually their reading of scores on tests involving correction formulas will be exactly right, although one point margin of error in the scoring of correction-formula tests is an allowable standard of accuracy in most machine-scoring units. Where $S = R$, the reading should be exactly right in every case.

An operator should know the principle on which the machine operates, the procedure for setting the machine, all scoring procedures, how to use the fields, how to prepare keys including the rights key and the item elimination key, how to insert keys in the scoring rack, how the controls operate, how to set the machine empirically for scoring by formula, when and how to use the cross-line circuit, how to check the machine for accuracy, how to do special operations such as the scoring of weighted item tests, how to use the aggregate-weighting unit, how to use the item counter, and the routine care of the machine. The operator need not be a skilled mechanic; in fact, not only is he not expected to correct mechanical faults in the machine, but he is definitely expected to call in the serviceman furnished by the company whenever the machine requires adjustment. Two machines permit scoring while one machine is out of commission. Service centers are located in the larger cities.

The International Test Scoring Machine does not represent the only machine-scoring method. Although it is unquestionably the one used almost exclusively at present, certain other machine procedures are of some interest historically, if not practically. About the time the electrical test-scoring machine became available, Cuff (112, 113) published a number of journal articles in which he described a novel scoring apparatus called a *noncalculator* which weighed the correct responses in a test and assigned scores based on weight. This machine is sold by Multiplex Display Fixture Co., St. Louis, Missouri. Its operation is wholly mechanical.

Uses of Special Equipment on the Scoring Machine

The uses of two kinds of special equipment in the IBM scoring machine should be explained briefly. These are the item counter and the aggregate-weighting unit.

The item counter

The item counter is an attachment on the International Test Scoring Machine which makes possible a graphic count of the responses given by a group of examinees to a test item. Ninety response spaces and up to 100 cases may be handled at one operation.

Not all scoring machines are equipped with item counters. They are installed only on special request and at an additional rental fee. As a rule, one item counter is sufficient for a machine-scoring unit, unless it is engaged primarily in research rather than scoring service.

The most important use of the item counter is in research, particularly item analysis. Its main use as a part of scoring service is to inform teachers concerning the success of their classes on each question in the test. This information will indicate areas of weakness and occasionally serve as a basis of corrective teaching. The item counter will record either the number of individuals getting the item right or the number choosing each response position. In research, information concerning the number of persons of a given criterion score choosing each response is often important, but in connection with scoring service it is generally sufficient to record only the number answering each question correctly.

Item counting is a process separate and distinct from scoring and is ordinarily done after the papers have been scored. Carbon paper is used with the counter to record the responses graphically on a standard form. An illustrative record is shown in Figure 15.

Aggregate-weighting unit

The aggregate-weighting unit may be used to calculate the weighted average of test scores, school grades, personality ratings resulting from oral interviews, and other measures. It makes possible the weighted averaging of as many as thirty measures with values of from 1 to 100, each of which is to be weighted from 1 to 20. If the values should vary from only 1 to 5 units, as many as sixty measures could be averaged. Thus, it solves the multiple regression equation from scores of the individual checked on the special weighting sheets, one of which is used for each person.

The aggregate-weighting unit, after having been properly plugged to give the different desired weights, is fitted into the place in the machine

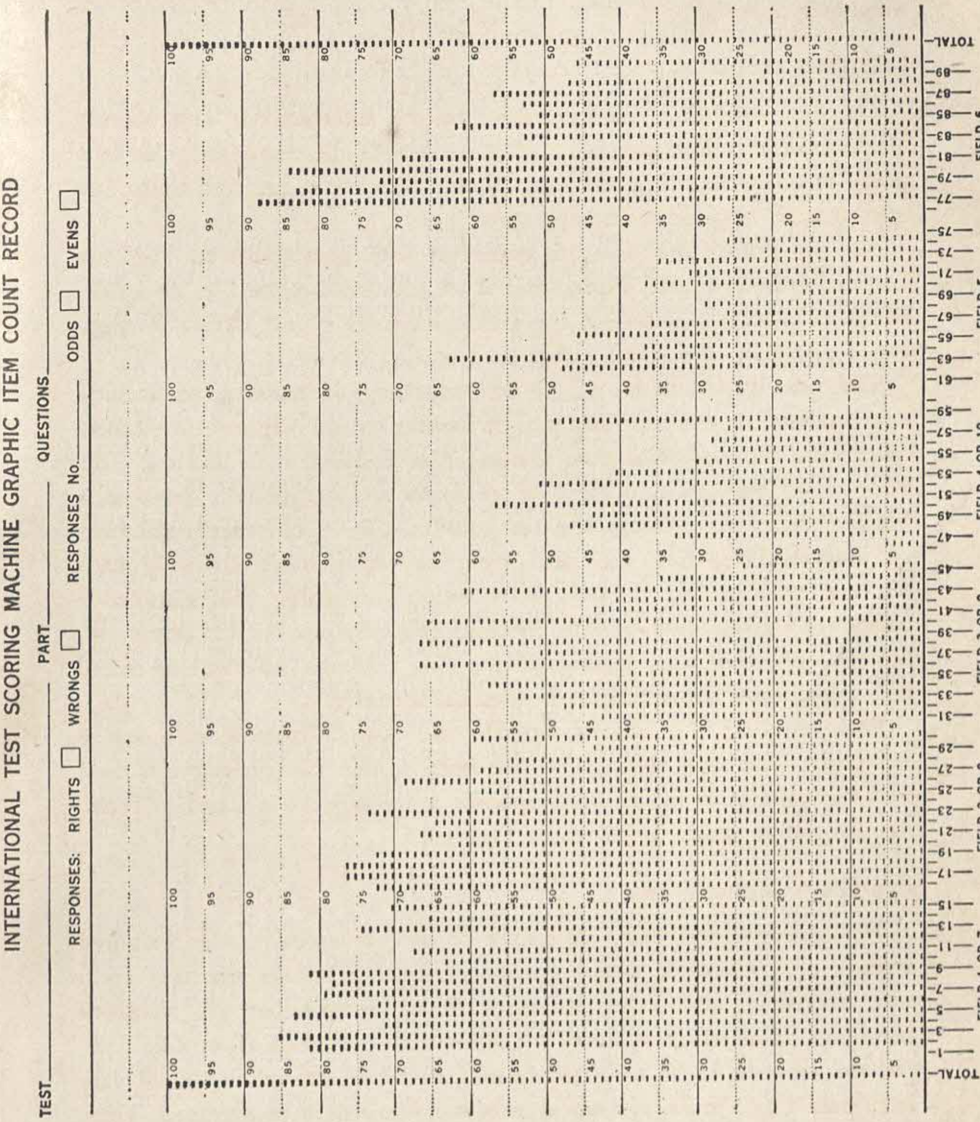


FIG. 15.—Record form for item counting. (Reduced about one-third.)

usually occupied by the scoring rack. The different values to be weighted and averaged are recorded with continuous lines drawn with an electrographic pencil on standard aggregate-weighting sheets. These sheets are then put through the scoring machine in the same manner as answer sheets. A sample aggregate-weighting sheet is shown in Fig. 16 (*41*, p. 13). The score on Daily Test 1A is 83; the one on Daily Test 2A is 82; and so on. The weights assigned to these tests and other measures are, of course, not shown on the sheet. They are determined by the way the unit is plugged.

Uses of Punch Card Equipment in Scoring

Although the International Test Scoring Machine is the main vehicle for scoring tests mechanically, various authorities have shown the values of punch card and mark sense equipment in scoring.

Recommended Procedure for Small-Scale Manual Scoring

Notwithstanding the availability of scoring machines in a considerable number of schools and colleges and in various service and research bureaus, a large proportion of the scoring of objective tests continues to be done by hand. Some manual-scoring projects include only a few test papers in a single school, while others involve many thousands and even hundreds of thousands of tests from many different schools. As here defined, small-scale manual scoring refers to the scoring done locally in one school, and large-scale scoring projects are those carried on by research departments of city school systems, by the central offices in state testing programs, and by test service organizations which are more or less national in scope. There are certain similarities in all scoring programs, but there are also differences in organization and procedures which depend upon the size of the project.

Scoring by teachers versus scoring by clerks

When the scoring is done locally, one of the first and most important questions concerns the personnel to be used in this work. The most common but least satisfactory use of personnel for this purpose is to have the teachers or counselors do the scoring. Staff members who have full teaching or counseling loads and who are given the work of scoring as an added duty can hardly be expected to be as efficient or accurate as a group of trained clerks. Moreover, the routine drudgery of scoring takes valuable time and energy which should be used in planning and carrying on instruction and guidance. It is sometimes assumed by administrators and school boards that the comparing of a pupil's objective responses with an answer key will, by means of some mysterious process, give a teacher or a coun-

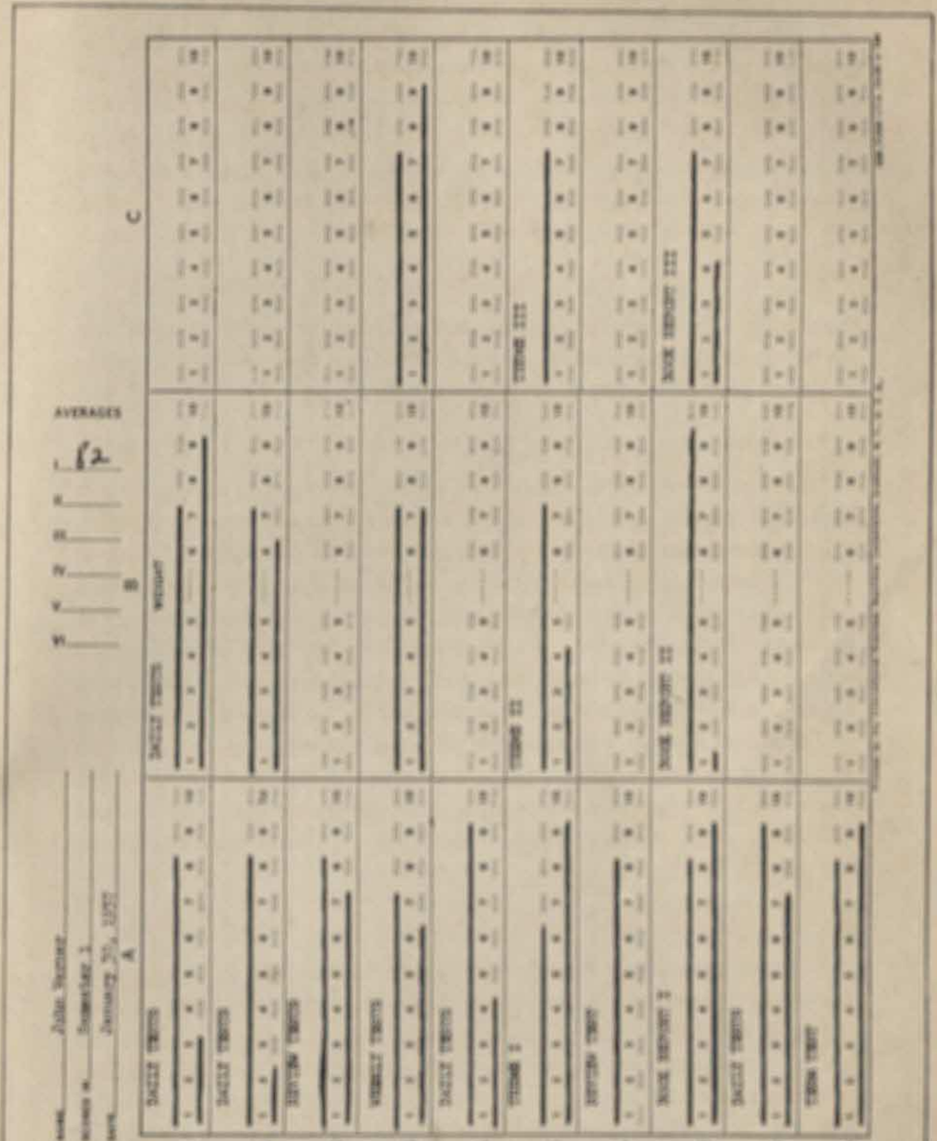


FIG. 16.—Sample aggregate-weighting sheet from *Manual of Instruction for the IBM Test Scoring Machine.* Reproduced by permission of International Business Machines Corporation. (Reduced about one-fourth.)

selor a diagnostic insight into the pupil's strengths and weaknesses. This could be a valid assumption only if a scorer of objective tests were obliged to read each question in order to evaluate each answer, but efficient scoring is not done in that way. A scorer compares mechanically and as rapidly as possible a key with the answers of the pupils and checks or counts the right and the wrong responses in routine fashion. The more "mechanical," the better. When the responses are entered on separate answer sheets, the scorer does not even have the test questions at hand. If the responses are recorded in test booklets, it is true that a teacher *could* take the time to study each pupil's paper while scoring it, but this procedure would diagnose rather than score. Diagnosis is a necessary step in the interpretation of test results, but scoring and diagnosis are essentially different procedures, and each one can be carried on more efficiently if the two processes are fully divorced. Moreover, if the school pays its teachers more than it pays its clerks, it is financially advantageous to have clerks do the clerical work and to free teachers and counselors for more productive activities on a higher professional level.

Training of scorers and procedures for eliminating errors

If a school adopts the practice of hiring full-time scoring clerks on a temporary basis following each testing program, it is probable that in most local situations from two to six clerks can score and tabulate the results of all the tests within a period of two or three weeks. It is realized, however, that until those who control the purse strings of the schools recognize the unwise economy in the use of teachers or counselors for scoring, the instructional and counseling staff will continue to do this kind of work because no other services are available for it. Since scoring is, under these circumstances, a part-time and after-school job, it may be necessary for all or nearly all the staff members to participate in the project in order to complete it within a reasonable time. Thus, the number of different individuals who will need to be trained as scorers may be fairly large. It is desirable to observe the following training rules:

1. Assign each individual to work on just one test, and make sure that he understands fully how the scoring is to be done.
2. Rescore the first ten papers scored by each individual, and make suggestions for the elimination of errors as needed. If errors are found, rescore the first twenty after correction of the faulty procedure.
3. If the responses have been entered on separate answer sheets, continue to have all papers rescored, using the less experienced individuals for the first scoring, and the more experienced ones for the rescoring. Check all discrepancies between scorers. Make systematic personal reports on errors to individual scorers.

4. If the responses have been entered in test booklets, it usually is not feasible to rescore completely the booklets for all pupils, once the accuracy of individual scorers has been found to be satisfactory. Check the scoring of every fifth paper to make sure that systematic errors are not occurring in the work of some individuals and check the addition of part scores on all papers. Employ the known dependable scorers for this function.

5. Completely rescore the test papers marked by all individuals who average more than one scoring error a paper, or, better, transfer such people to other work or discontinue their service.

Scoring by pupils

Reference was made in an earlier section to a plan used in connection with the Iowa Every-Pupil Tests of Basic Skills for reducing the amount of clerical work in scoring by having the pupils themselves mark the omitted items. A question may appropriately be raised concerning the feasibility of going further and of having the pupils completely score their own papers under supervision. The question is especially pertinent when the tests are achievement tests and when it is planned to use them as teaching devices. Some experimental evidence favorable to the value of pupil scoring of test papers was obtained by Curtis and Woods (*16*), who compared the learning value of three ways of teaching correction of new-type tests with a plan of having the pupils score the papers and concluded that "when new-type examinations are made a teaching device, the method of correction which requires the least use of the teacher's time and energy, namely, that under which the pupils check the incorrect items on their own papers during a discussion of the test items, is the most profitable for the pupils of the four methods studied."

It is obvious that it is feasible to consider the idea of pupil scoring only if that type of scoring will be done honestly and with reasonable accuracy. Hoff (*33*) found that pupil scoring of objective-type examination papers could be done as accurately or more accurately than teacher scoring if the work was controlled by having the pupils exchange papers with individuals other than friends, and then having the "checkers" cross out the omitted items and sign their names before scoring. When pupils scored their own papers, the errors averaged 7–8 percent, and thus the results were too unreliable for accurate grading.

Well-planned test scoring by pupils might be feasible if all teachers were careful and accurate scorers and were capable of supervising the scoring procedures. The only plan that would be workable from an administrative point of view would be to have the pupils do the scoring immediately after the tests were given and in the same groups in which they were taken. The same teachers who administered the tests would,

as a rule, serve as scoring supervisors. The idea of having inexperienced pupils score tests under the supervision of teachers who are themselves inexperienced scorers is a hazardous one at best, and should be employed only as a last resort.

There is apparently very little relationship between intelligence and accuracy of scoring. Studies by Pintner (*43*), Dearborn and Smith (*18*), Herbst (*30*), Rauth (*46*), Dunlap (*23*), Traxler (*62*), and others indicated that teachers and college students preparing for teaching are not naturally accurate scorers. Rauth reported data on the accuracy of scoring several intelligence and achievement tests. The general indication was one of considerable inaccuracy in scoring. Dunlap studied teachers' errors in the scoring of the subtests of the Terman Group Test of Mental Ability. Tests particularly likely to be misscored were the true-false, yes-no, and the same-opposite tests, where the subject was required to underline one of the terms, and tests where the subject underlined two words in the test. Traxler found that teachers and graduate students did not score semiobjective tests, such as certain sections of the Metropolitan Achievement Tests, accurately on a scorers' test designed to bring out certain pitfalls. The mean percentage of items scored incorrectly was 18.4. The fast scorers were as accurate as the slower ones.

The standards of accuracy in small-scale local scoring, where the part-time services of teachers or others are used, probably cannot be expected to be as high as the standards for large-scale programs. Greater accuracy can be achieved in the scoring of separate answer sheets than in booklet scoring because all operations in answer-sheet scoring can conveniently and economically be done twice. Experience in supervising the local scoring of test booklets for a school over a considerable period indicates that an error average of one point per test paper is a satisfactory and realistic standard for a scoring program of that kind. The electrical and mechanical test-scoring machines should show a trifle better average record than that.

Provision for organization and tabulation of the results

In addition to the scoring clerks, a small-scale scoring staff should include at least one or two individuals who are especially trained to make distributions, do simple statistical work, translate raw scores into standard scores, and enter percentile ratings on the papers. The changing of raw scores to standard scores is an especially important step, for in this step opportunities for very large errors occur, unless the work is done very carefully and is checked meticulously.

As soon as the tests have been scored, distributions of scores should

be made, class medians should be found, percentile ratings should be reported for the scores of the individual pupils, and typed lists of scores and percentiles should be prepared. Copies of these lists should be distributed to the teachers who are concerned with the different class groups and subjects tested. It is also desirable to enter the scores either on cumulative records or on individual profile sheets which are especially prepared to cover all the tests at a given class level in one testing program and which may be filed by the counselors in individual pupil folders. The test results should be explained by the scoring department to the teachers and counselors in individual and group conferences.

Organization of Large-Scale Manual-Scoring Procedures

A large-scale manual-scoring project requires the close cooperation of many individuals performing different tasks. The procedures must be thoroughly organized and outlined in great detail. The functioning of the program must constantly be under the watchful eye of a general supervisor who will maintain a steady flow of test materials from department to department and who will be alert to the need for anticipating and avoiding bottlenecks which can throw the whole program out of gear just as they can in any industrial plant working on an assembly-line basis.

Departmental organization

The departmental organization will depend partly upon whether the testing program is a completely prescribed one or is one in which local schools were allowed to select tests from a suggested list. If the program is prescribed, and if the administration has been so arranged that the results for all the tests in a battery have been entered on a single answer folder, the actual scoring can ordinarily be handled in one large department. If the program is quite varied and involves a large number of different kinds of test booklets, several scoring departments will be needed. For example, in connection with the spring testing program of the Educational Records Bureau there are, in addition to the machine-scoring department, manual-scoring departments for tests in the elementary school subjects, secondary school English, secondary school mathematics and science, and secondary school foreign languages and social studies. Departments needed in addition to those for scoring include a department for the receipt and classification of the tests, a department for the routine reporting procedures—either statistical typing or punch card and tabulating machines—a department for assembling report lists for each school and checking the reported scores against those on the booklets or answer

sheets, and a department for the preparation of explanatory and interpretative letters to accompany the reports which will be returned to the different participating schools.

Selection of scorers

Since testing programs tend to be seasonal, a large number of temporary scoring clerks needs to be added to the staff for periods of a few weeks at different times during the year. The selection of scorers should be made on the basis of an interview and one or more tests. Experience indicates that the type of tests for scorers that has the highest predictive value is a work-sample test in which a regular scoring situation is approximated as nearly as possible. Many copies of the same standardized test filled out as if the students had taken them are prepared. The applicants, who may be tested in groups if the number is large, are supplied with copies of the test and scoring keys. They proceed to score the test according to a definite set of directions. Their scoring is then evaluated for accuracy and speed. Under normal employment conditions, it is advisable to consider for scoring positions only those individuals whose percentiles, based on distributions of the scores of the applicants, are above 75 in both accuracy and speed. A page from a test used in the selection of scorers of semiobjective tests at the Educational Records Bureau and the corresponding portion of the scoring key are shown in Fig. 17.

Routinization

In the scoring of tests on a large scale in connection with a prescribed program, the functionaries needed in addition to the supervisors are a receiving clerk, a counting clerk, a file clerk, sorting clerks, scanning clerks, scorers, checkers or rescorers, referees, distributors, and typists or tabulating-machine operators.

The receiving clerk receives and opens the packages and gets the materials ready for the counting clerk. The first counting clerk counts the used and unused answer sheets returned as well as the test booklets returned and checks against the supervisor's report. The second counting clerk recounts the used answer sheets and checks any irregularities found by the first counting clerk.

The file clerk transfers data from the supervisor's report to a card file for the schools participating in the program. He also keeps a file of the supervisor's reports. The sorting clerks sort the papers into appropriate age, sex, class, or other groups. They then separate the papers into convenient scoring units of perhaps fifty papers each and attach an identification label to each group. The scanning clerks check the number of papers in each

ENGLISH: PART I—LANGUAGE USAGE

Directions. In each sentence one word is left out for each blank line. Think of the one word that should be written on the blank line to make the sentence correct and *sensible.* Write the word in the parentheses after the sentence. Read the whole sentence before you write the word. (Sometimes the first letter of the word you are to use is given.) *Read the sentence again* after you have written the word, to be sure that it is correct and sensible.

Samples. John says he *d*doesn't like to lose a game .... (*d*oesn't)
Mary hasn't eney more paper left for her lessons. (eney)

			~~doesn't~~
			~~any~~
1.	Chester had hardly eney marbles when he started to play..	(eney)	any; or enough
2.	My aunt invited me for a week, but I *st*adyed only two days.	(*st* adyed)	stayed
3.	My father has *g*even much money to the Red Cross......	(*g* geven)	given
4.	The last boy *w*how we invited to the party cannot come.	(*w* whow)	whom
5.	*W*—— you surprised to see us?	(*W* where)	were(n't)
6.	The entire audience *w*—— interested in the speech	(*w* as)	was
7.	The swarm of bees followed —— own queen until they reached the new hive..............................	(there)	their
8.	Mother said to Blanche, "I *br*—— you a very nice present."	(*br* ut)	brought
9.	*W*—— do you think will win the prize?................	(*W*how)	who
10.	Yesterday we learned about Boston. Today Miss Johnson —— us a lesson about New York	(tort)	taught; or gave
11.	If Dick hadn't *thr*owen the ball away, we would still have been playing..............................	(*thr*owen)	thrown
12.	"Where is Grace?" "She has goan to Judith's for lunch."	( )	gone; or been
13.	Everyone wanted his own seat..	(his)	her; or his
14.	How soon must we go? It is nerly time to go now.......	(nerly)	almost; or nearly
15.	Your book has been *l*ay*ing* on the table since morning.	(*l* aying)	lying
16.	Ned would not be so untidy if he could only see him*self*..	( )	him(self)
17.	We had to walk very *qu*—— in order to get there on time..	(*qu*ickle)	quickly
18.	Millie *h*—— herself when she fell downstairs............	(*h* eart)	hurt
19.	Of the three sisters, Peggy is the *pr*——	(*pr*oudest)	prettiest (*any superlative*)
20.	I don't care for most apples, but I do like those *k*—— of apples..............................	(*k*inds)	kinds
21.	The council —— its village mayor every two years......	(select)	elects; or chooses (*any suitable singular verb*)
22.	Paul, you *c*—— late to school every day last week.......	(*c*ame)	came
23.	If Dan had looked for the skate, he *w*—— —— found it.	(*w*ould of)	would have
24.	I recognized John as soon as he came *i*—— the room.....	(*i*nside)	inside; or into
25.	"I have already *ch*—— Agnes for my partner," said Mary.	(*ch*ose)	chosen
26.	Fred is just beginning, but Jenny *b*—— ten minutes ago..	(*b*egain)	began
27.	To *w*—— are you going to send that pretty valentine?...	(*w*om)	whom
28.	I didn't say ——*thing* to Miss Smith about the broken vase.	(eney)	any(thing)

ENGLISH PART I

SCORING KEY

[1] *(Go right on to the next page.)* (1)

FIG. 17.—Portion of test and key used in the selection of scorers of semiobjective tests at the Educational Records Bureau. Two separate forms are combined here.

scoring unit, scan the papers for irregularities not noted by preceding clerks, and erase or mark out multiple answers to any questions.

Assuming that answer sheets rather than test booklets would be used in a large-scale uniform program of this kind, the scorers count the number of correct answers (and also the number of incorrect answers, if a correction formula is being used), obtain the score and enter it in a designated place. The rescorers score the papers a second time, preferably using a scoring key devised in such a way that they cannot see the scores obtained by the first scorers. A third person, or "referee," who should be an especially able and accurate scorer, goes through all the papers, selects those on which there are discrepancies between the scorer and rescorer results, and scores these papers again in order to decide what score is correct.

The distributors make distributions of the scores according to the classes, schools, and other groups as needed. The statistical typists or tabulating-machine operators prepare copies of the data in a form suitable for returning to the participating schools. The checkers get everything for a given school ready, check the data against the original sheets, and compare numbers with those given on the supervisor's report or inventory sheet. The results are then ready to be returned to the participants.

The routinization in connection with a program in which each local school chooses from a suggested list of tests those it desires to use is somewhat similar to that described above, but it cannot be as complete and as detailed. When the tests are received from the schools they go to the classification department where they are inspected for discrepancies and carefully separated into the proper groups as determined by number of years of study and other factors. It usually is not feasible to put the papers into scoring units of a definite number of papers each. Those for a given school are customarily scored as a unit regardless of how many papers each pack contains. The papers are taken from the classification department into the various scoring departments, and the work is so planned that the scoring of all the tests for a given school will be completed about the same time. As the scoring of the test booklets for a school is finished, distributions of the scores are made, and medians, quartiles, and other needed measures are computed.

The test booklets go to the typists who type alphabetical class lists showing the scores and percentile ratings of the various pupils. If IBM cards have been prepared for each pupil, this can be done automatically on the tabulator. These lists may be made in triplicate or quadruplicate. When all distributions and lists have been made and the work has been checked, the various parts of the report are assembled and inspected by the chief

scoring supervisor. The assembled report, together with the school's statistical folders for preceding testing programs, is sent to the desk of a staff member who is trained in psychological measurement and guidance techniques. He dictates a covering report explaining and interpreting the scores for the school to which the report is being sent.

The use of tabulating equipment has not been found to be feasible in making reports on tests in a program which varies from school to school.

The scoring schedule

The planning and carrying-out of a scoring schedule involving tests from several hundred schools, as some large-scale scoring programs do, is a fairly complicated procedure, and it calls for the full cooperation of all the supervisors in the scoring unit. It is desirable to schedule the work so that the report of the results of tests for a school will be returned within about two weeks after the test papers are received, and provision should be made for much faster service in cases where the results are urgently needed.

It has been found desirable to have a meeting of the supervisors of all the departments in the scoring program each week, at which time the schedule is prepared for the reporting of the results of all tests received since the last meeting. The schedule is arranged by dates and by schools under each date. Typed copies of each week's schedule are made, and all supervisors concerned are provided with a copy. It is then the responsibility of each supervisor to plan and keep up the flow of work in her department, so that the schedule can be maintained. The head supervisor keeps in constant touch with the departmental supervisors and helps meet emergencies in different departments through shifting scoring personnel and other adjustments. Each school's tests as they come in are stamped or otherwise designated with a job serial number. These serial numbers are helpful to the supervisors in organizing their work so that the scoring of all the tests for any one school will be finished at approximately the same time.

Forms used

Numerous forms are employed in a scoring project, but perhaps the most important one in connection with the routine of the actual handling of the papers is a classification slip. This slip shows all pertinent data relative to the source and composition of the papers and indicates each step taken in the classification, scoring, and reporting procedure, including the initials of *all* workers who handle any phase of the work and the person (initials) responsible therefor. The slip also provides a detailed

NO. 96 NO. PAPERS 23 GRADE 9

SCHOOL North East Jr. High School YR.OF STUDY

Berkeley, Mo. LOCAL COURSE

SUBJECT English Mr. Lang

TEST Cooperative A, B1, C1

FORM T

CLASSIFY EK CHECK JB HOLD,DISC. ABSENTEES

	SCORE			RESCORE				COUNT			RECOUNT			
PART	INIT.	IN	OUT	INIT.	IN	OUT	ERRORS	INIT.	IN	OUT	INIT.	IN	OUT	ERRORS
1 A	TM	9^{20}	10^{45}	BL	2^{00}	2^{35}	0/5				AD	12:13	12:30	0/17
2 B	TM	11^{00}	11^{45}	BL	2^{35}	2^{55}	0/5				AD	1:35	1:50	0/17
3 C	TM	12^{00}	1^{45}	BL	3^{00}	3^{20}	1/10				AD	2.30	3.00	1/12
4														
5														
6														
7														
8														
9														
10														

				CHECK							CHECK		
OPERATION	INIT.	IN	OUT	INIT.	IN	OUT	OPERATION	INIT.	IN	OUT	INIT.	IN	OUT
Chron.Age							Alphabetize	CS			HW		
Scan							Match						
Mark							Profile						
Transfer	mp	10:15	10:46	KP	1^{15}	1^{31}	Plot						
Add							Connect						
Sc. S. Parts							Class Prof.						
Sc. S. Total							Cl. Md. Sheet						
Recheck							Cl. Anal. Ch.						
Average							Distribute	CS	3:30	3.40	HW	3^{45}	3^{55}
Mental Age							Part Scores						
I. Q.							Total						
Educ. Age							Q-Score						
E. Q.							L-Score						
A. Q.							C. A.						
P. S. Grade							M. A.						
Av. Grade							I. Q.						
Acad. Apt.							Acad. Apt.						
Decile													
Total %ile													
Part %ile							Type	an	4^{00}	4^{15}	CS-MT		
Local %ile													
Item Anal.													
Over-print.													

FIG. 18.—Specimen classification slip for recording routine of handling the scoring of a test. Also provides a detailed record of scoring errors.

record of scoring errors. A specimen classification slip is shown in Figure 18.

Motivation of scoring clerks

A noteworthy problem in the rapid and efficient handling of a large-scale national scoring program is to motivate the scorers to put forth their best efforts. In a program which is not uniform from school to school nor from test to test, it is hardly feasible to set up routine procedures for motivation. Judicious praise and increased pay for superior performers are the main ways of motivating scorers in this situation.

In a uniform testing program which can be highly routinized, however, a simple scheme for motivation based upon scoring units of fifty papers each has been found to give good results. The motivating factor consists in paying a scorer a bonus (usually an extra five or ten cents) for each pack of fifty papers scored without error. While the amount is small, it gives the scorers something definite for which to work, sometimes stimulates friendly rivalry among them, and helps to build up their morale when they have done well. This kind of motivation encourages both accuracy and speed of scoring.

Standards of accuracy

The accuracy standard which can be maintained depends upon the kind of test, its length, the type of response required, the way of reporting the response, the number and quality of the scorers available for employment, and other factors. For example, the Metropolitan Achievement Tests, which are long and contain semiobjective questions in certain parts, are far harder to score than the American Council Psychological Examination in which the questions are entirely objective and for which the responses are recorded on separate answer sheets. Even with the greatest of care, an occasional error will creep into the scoring of test booklets, but there should not be an average of more than one one-point error in five papers, and there should be no error large enough to change the percentiles significantly.

In the scoring of answer sheets used in a prescribed testing program, there should be almost no errors in scores released by the scoring department. In the scoring of the Army-Navy Qualifying Tests, one of the most extensive uniform testing programs ever carried on, the scoring and checking procedures used made errors in reported scores highly unlikely, and the routine of scoring was simplified to such an extent that the best of the individual scorers made some almost phenomenal accuracy records. An occasional scorer did not average more than one error in 500 papers

each containing 160 items, and even the poorest scorers retained on the project did not make more than one error in about 25 papers.

Summarizing the results

At the end of a large-scale scoring project, the results for all the participating institutions should be collated and summarized. Norms should be prepared if suitable ones are not already available. The results should be printed in a form which will help each school use the data in self-appraisal. A sample chart of this type is shown in Fig. 19.

Manual versus Machine Scoring

Organizations providing scoring services and schools large enough to use machine scoring need to decide whether the scoring should be done manually, mechanically, or through a combination of these two procedures. The decision will depend largely upon relative cost, comparative accuracy, and relative usefulness of the test results.

Data in the files of test service organizations show without question that machine scoring of *answer sheets* is considerably cheaper than manual scoring of *test booklets.* There is also evidence that it is less expensive than manual scoring of answer sheets in connection with a testing program which is relatively flexible.

Machine scoring may not, however, be faster or cheaper than manual scoring of a prescribed uniform program where all manual scoring procedures can be highly routinized.

Most of the comparisons thus far made between manual and machine scoring have been poorly controlled, and the differences in favor of the machine could often have been attributed to other factors than those in the use of the machine itself. When machine-scoring methods are compared with primitive or badly designed manual-scoring methods, the machine naturally proves markedly superior. When sufficient care is taken in the design of answer sheets, scoring keys, and scoring routines, however, so that maximum efficiency in both manual and machine scoring is obtained, it is by no means certain that machine methods are any better than manual-scoring methods in most situations.

One example of the possibilities of carefully planned manual scoring is provided by the Iowa Testing Programs, in which more than three million test scores are obtained annually by manual-scoring methods. The answer sheets, scoring key, and procedures used in scoring the Iowa Tests of Educational Development have been illustrated earlier in this chapter (pages 379–82). Lindquist reports that in the long run his better scorers average more than 325 scores per hour on tests which average 80

CONFIDENTIAL SUMMARY REPORT

of Results on the

IOWA TESTS OF EDUCATIONAL DEVELOPMENT

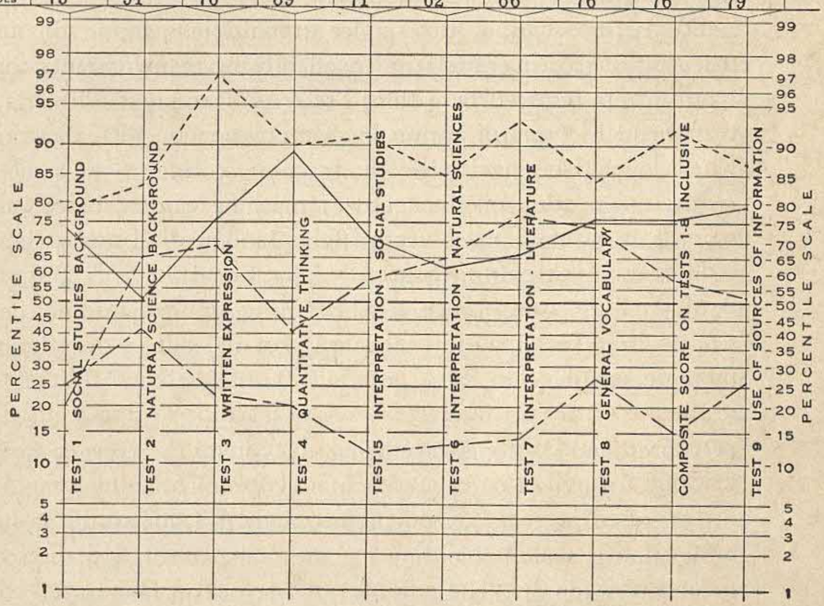
AVERAGE STANDARD SCORES FOR EACH GRADE SEPARATELY, WITH THEIR CORRESPONDING PERCENTILE RANKS AMONG SCHOOLS IN THE SAME ENROLLMENT CLASS AND AMONG SCHOOLS OF ALL SIZES

For the Midtown High School Y B Enrollment Class

(Class A=275 or more *tested* in grades 10-12 incl.; B=91-274; C=46-90; D=45 or less)

GRADE / NUMBER TESTED		TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6	TEST 7	TEST 8	TESTS 1-8	TEST 9
GRADE 9	AVERAGE STANDARD SCORE	10.3	11.0	10.9	09.5	08.8	09.0	08.7	10.4	09.2	09.4
NUMBER TESTED 49	PERCENTILE RANK IN ENROLLMENT CLASS	12	35	09	09	01	04	03	09	03	15
	PERCENTILE RANK AMONG ALL SCHOOLS	25	42	23	20	12	13	14	27	15	26
GRADE 10	AVERAGE STANDARD SCORE	11.2	14.9	14.2	12.7	12.8	13.2	13.5	14.1	13.6	13.0
NUMBER TESTED 46	PERCENTILE RANK IN ENROLLMENT CLASS	11	56	62	29	51	60	74	68	48	37
	PERCENTILE RANK AMONG ALL SCHOOLS	20	65	68	41	58	65	77	74	58	52
GRADE 11	AVERAGE STANDARD SCORE	15.9	16.9	18.3	16.9	16.6	16.0	16.4	16.5	17.5	17.6
NUMBER TESTED 47	PERCENTILE RANK IN ENROLLMENT CLASS	75	82	99	93	92	88	94	82	94	87
	PERCENTILE RANK AMONG ALL SCHOOLS	78	83	97	90	91	85	93	84	92	87
GRADE 12	AVERAGE STANDARD SCORE	18.1	16.1	17.4	18.0	17.0	16.1	16.2	17.5	18.0	18.5
NUMBER TESTED 36	PERCENTILE RANK IN ENROLLMENT CLASS	68	42	70	92	61	49	59	72	71	73
	PERCENTILE RANK AMONG ALL SCHOOLS	73	51	76	89	71	62	66	76	76	79

PROFILE
of
PERCENTILE RANKS
of
GRADE AVERAGES



LEGEND: ——————— Grade 9 – – – – – – – Grade 11
— — — — Grade 10 ———————— Grade 12

SEE OTHER SIDE FOR EXPLANATION OF PROFILE

FIG. 19.—Sample of summary chart collating results from all schools participating in Iowa Tests of Educational Development.

items in length. One of his scorers averaged over 450 scores per hour over a period of several days. He states that for the total of 69 scorers employed in 1948–49, including the results from some who were tried out and released after a week's trial, the year-round average was 279 scores per hour per scorer. These figures are based on the total employed time of the scorers, including rest periods, time spent in "make-ready" and so forth. In this program, nine tests were scored on a single answer sheet.

With machine scoring, five answer sheets (nine sides) and nine insertions in the machine would have been required to score the same battery for a single pupil. With a small change in the numbers of items in the various parts of the test, however, the battery could have been made to fit upon three answer sheets (six sides) and the number of insertions in the machine reduced to six.

The manual-scoring procedure also eliminates the cost and inconvenience involved in the use of special electrographic pencils, as well as the remarking of poorly marked sheets. In the Iowa program, each paper is scanned before scoring for double or multiple marking of items, and each sheet is scored entirely independently by two different scorers, and disagreements resolved by a third "referee." Lindquist estimates that the over-all cost of manual scoring in that situation is only a fraction of the cost that would be involved in the machine scoring of the same tests.

Some information on the speed of machine scoring, based on less extensive records than those reported by Lindquist, is available from the machine-scoring department at the Educational Records Bureau. These figures are derived from relatively brief runs, and it should not be inferred that the same speed would necessarily be maintained over long periods of time comparable to those on which Lindquist based his hand-scoring data.

Q-scores and L-scores, averaging 100 items a score, were obtained from the American Council Psychological Examination by a rapid machine operator of many years' experience at a rate of 1,004 scores an hour. This examination is scored according to the number right, and the scoring can be done very rapidly. The rate of 1,004 scores an hour is an average for scoring and rescoring. The answer sheets were not scanned and re-marked before scoring. Papers on which there were discrepancies between the two readings were scored by hand. The hand-scoring time, when added to the machine-scoring time, reduced the average rate to 690 scores an hour.

The Cooperative English Test, Test A, Mechanics of Expression, which yields one score based on 159 items, corrected for guessing, was scored by an experienced machine operator at the rate of 536 scores an hour.

These answer sheets, like those for the American Council Psychological Examination, had not been scanned and re-marked before scoring. When the time for hand-scoring discrepancies between the two readings was included, the average was reduced to 361 scores an hour.

The machine-scoring time on these two tests is on the whole somewhat more favorable than the time reported by Lindquist for his exceptionally efficient hand-scoring organization. There is, however, an important limitation in the fact that when answer sheets are scored without scanning, one cannot be sure of the accuracy. While a poorly marked paper will usually result in a discrepancy between the two readings of the first and second scorings, if the answer sheets are scored in one machine and rescored in another, one cannot be sure that there will be a difference in readings if the machines are equally sensitive.[2] In order to insure correct readings, it is necessary to scan the answer sheets and re-mark those that contain light marks as well as to erase stray dots.

The time required for scanning poorly marked answer sheets greatly reduces the output. For instance, the Iowa Silent Reading Tests, Parts I and III, yielding three scores averaging 25 items per score, was scored, after scanning, at a rate of 775 scores an hour. But the scanning time, when added to the machine-scoring time, reduced the rate to 283 scores an hour. Similarly, the Cooperative English Test, Test B2, Effectiveness of Expression, which contains 65 items and is scored with a correction formula, was machine-scored, after scanning, at a speed of 472 scores an hour. The scanning time, however, brought the speed down to 100 scores an hour. It appears from records kept for a week in the machine-scoring department of the Educational Records Bureau involving much scanning of groups of very poorly marked answer sheets that the average output for the week was not more than 100 scores an hour.

On the other hand, when answer sheets are well marked so that the scanning time is reduced to fifteen seconds a paper, or less, machine-scoring output seems to compare favorably with even the best hand-scoring output thus far reported. For instance, in connection with one group of well-marked answer sheets for the American Council Psychological Examination, the speed, including scanning, was 468 scores an hour.

These data suggest that the actual speed of machine scoring is considerably faster than even the most rapid hand scoring carried on under optimum conditions. However, if a large amount of scanning and re-

[2] Some interesting experimentation along this line has recently been carried on at the Educational Testing Service. One scoring machine has been "doctored" to reduce its sensitivity. Unscanned answer sheets are scored in an ordinary machine and rescored in this less sensitive machine. This procedure reduces the chance of getting the same incorrect reading due to poor marking of the papers in the two machines.

marking of answer sheets is necessary before they are run through the machine, the net result in number of papers processed is apparently not as good under reasonably efficient machine-scoring conditions as it is in a very well-planned and administered manual-scoring program. Superiority of one method over the other apparently depends upon the question of the amount of scanning necessary. Under present conditions, as indicated earlier in this chapter, most groups of answer sheets received from schools and colleges are so indifferently marked that a large amount of re-marking is required as a preliminary to machine scoring. Machine scoring will not be superior to efficient hand scoring of answer sheets until test authors and examiners cooperate in seeing that answer sheets are marked so well by students that they will require very little re-marking but can almost immediately be inserted into the machine.

Regardless of the problem of re-marking, however, there can be little doubt that where the scoring involves the use of a correction formula machine scoring is much faster and less expensive than even the best organized and most rapid hand scoring of *test booklets*. In machine scoring, the scoring and correction is done in one process, and the pointer on the dial shows the *corrected* score. In the manual scoring of answer sheets, it is likewise possible to obtain a close approximation of the corrected score in one operation.

In accuracy as in speed, neither kind of scoring is unequivocally superior to the other. If all answer sheets are put through the scoring machine twice, machine scoring is more accurate than hand scoring of test papers in which part of the scoring process is done twice, and a sample check routine is used with the rest. It is not, however, more accurate than hand scoring of answer sheets under a procedure in which all processes are done twice. For tests in which the score is the number right, manual scoring and machine scoring of answer sheets are probably equally accurate if the scoring is done twice. But for tests scored with a correction formula, the number of one-point errors will probably be the greater on machine scoring.

The preceding discussion is concerned only with the relative *output per* worker under efficiently organized machine- and manual-scoring procedures. In determining the relative *cost* of the two procedures, one must also take into consideration, of course, the rental of the scoring machine, the considerably higher cost of machine scorable answer sheets, the possibility that more machine-scorable than manually scorable answer sheets may have to be employed, the cost of collating a number of separate answer sheets for the same examinee, the cost in both time and money of supplying, distributing, and collecting special electrographic pencils

for the examinees, the higher wages that ordinarily must be paid to machine operators than to scoring clerks, and the slightly greater amount of floor space that may be required for manual scoring.

As pointed out in an earlier section, some teachers feel that the results of manually scored test booklets are more useful than the results of machine-scored tests because the scored booklets themselves constitute a kind of diagnostic instrument which the teacher can study. There is no difference in the relative usefulness of hand-scored and machine-scored answer sheets. As indicated earlier, the key can be overprinted on either hand-scored or machine-scored answer sheets, and thus the correct and incorrect responses to the different items can readily be identified by the teacher or by the individual pupils, if the answer sheets are returned to them for purposes of class discussion and corrected teaching.

Many more answer sheets than is now the case could profitably be printed on cardboard for filing. This would facilitate their scoring as well. If the answers occur on one side of the sheet only, a student personnel form might be printed on the other. Then, for research purposes, cases could be sorted with respect to the personnel data, and item analysis might be done by the groups resulting from the other side of the form.

Appraisal of Current Tendency To Use Separate Answer Sheets

The growing tendency to employ separate answer sheets is perhaps the most pronounced single trend in objective testing. It unquestionably has a restrictive effect upon test construction, for it tends to force test items into a single-response pattern—the multiple-choice question. The use of other types of questions is not precluded, however, if test makers will use skill and ingenuity in setting up their answer sheets. More imagination and care than are typically employed in devising a test need to be used when setting up separate answer forms.

As indicated earlier in the chapter, there is not, at present, enough experimental evidence to warrant a definite statement concerning the effect of separate answer sheets upon the validity of the test results. It is appropriate to urge, however, that test authors have an inescapable obligation to find by means of research an answer to this question before they "go overboard" completely for answer sheet procedures.

The main justification for the use of separate answer sheets is that they are inexpensive, that they save time, and that they bring objective tests within the reach of hundreds of schools that could otherwise not use these tests at all. Through the continued availability of these inexpensive instruments, objective testing may eventually be brought to all schools in the United States.

Selected References

1. ANGELL, GEORGE W., and TROYER, MAURICE E. "A New Self-Scoring Test Device for Improving Instruction," *School and Society,* **67**: 84–85, Jan. 31, 1948.
2. BEDELL, RALPH. "Scoring Weighted Multiple Keyed Tests on the IBM Counting Sorter," *Psychometrika,* **5**: 195–201, September 1940.
3. BENNETT, GEORGE K. "Simplified Scoring Method for the Bernreuter Personality Inventory," *Journal of Applied Psychology,* **22**: 390–94, August 1938.
4. BICE, RAYMOND C. "More Effective Use of Machine Scored Examinations," *American Psychologist,* **2**: 179, May 1947.
5. BIRD, CHARLES. "The Detection of Cheating in Objective Examinations," *School and Society,* **25**: 261–62, Feb. 26, 1927.
6. BLOMMERS, PAUL, and LINDQUIST, E. F. "Rate of Comprehension of Reading: Its Measurement and Its Relation to Comprehension," *Journal of Educational Psychology,* **35**: 449–72, November 1944.
7. CASANOVA, T. "Weighting of Tests Measuring the Same Functions in Terms of Their Length," *Journal of Experimental Education,* **11**: 238–42, March 1943.
8. CHAPMAN, D. W. "The Scoring of Matching Tests with Unequal Series of Items," *Journal of Educational Psychology,* **27**: 368–70, 1936.
9. CHEN, L. "Correction Formula for Matching Tests," *Journal of Educational Psychology,* **35**: 565–66, December 1944.
10. CONRAD, H. S. "The Scoring of the Rearrangement Test," *Journal of Educational Psychology,* **27**: 241–52, April 1936.
11. COOK, WALTER WELLMAN. *The Measurement of General Spelling Ability Involving Controlled Comparison between Techniques.* ("University of Iowa Studies in Education," Vol. 6, No. 6, New Series No. 221.) Iowa City: University of Iowa, 1932. 112 pp.
12. COURTIS, S. A. *Why Children Succeed.* Delrest, Courtis Standard Tests, 1925.
13. CUFF, N. B. "A New Way to Score Tests," *Educational Method,* **14**: Part 2, 98–100, 1934.
14. ———. "Scoring Intelligence Tests by Weight," *Journal of Applied Psychology,* **20**: 769–77, December 1936.
15. CURETON, EDWARD E., and DUNLAP, JACK W. "Scoring the Rearrangement or Continuity Tests," *School Review,* **38**: 613–16, October 1930.
16. CURTIS, FRANCIS D., and WOODS, GERALD G. "A Study of the Relative Teaching Values of Four Common Practices in Correcting Examination Papers," *School Review,* **37**: 615–23, October 1929.
17. DAVIDSON, WILLIAM M., and CARROLL, JOHN B. "Speed and Level Components in Time-Limit Scores: A Factor Analysis," *Educational and Psychological Measurement,* **5**: 411–27, Winter 1945.
18. DEARBORN, W. F., and SMITH, C. W. "The Results of Rescoring Five Hundred Thirty Dearborn Tests," *Journal of Educational Psychology,* **20**: 177–83, 1929.
19. DICKENSON, H. F. "Identical Errors and Deception," *Journal of Educational Research,* **38**: 534–42, March 1945.
20. ———. "Pattern Sectioned Pupils in Lieu of Equivalent Test Forms for Classroom Testing," *Journal of Educational Research,* **33**: 183–88, November 1939.
21. DOUGLASS, H. R., and SPENCER, P. L. "Is It Necessary to Weight Exercises in Standard Tests?" *Journal of Educational Psychology,* **14**: 109–12, February 1923.
22. DUNLAP, J. W. "Problems Arising From the Use of a Separate Answer Sheet," *Journal of Psychology,* **10**: 3–48, July 1940.
23. ———. "Relationship Between Type of Question and Scoring Errors," *Journal of Experimental Education,* **6**: 376–79, March 1938.
24. ETOXINOD, COUNT SUSSICRAN (pseud.). "How to Checkmate Certain Vicious Consequences of True-False Tests," *Education,* **61**: 223–27, December 1940.
25. FEDER, DANIEL D. "Effect of Directions and Arrangement of Items on Student Performance in a Test," *Journal of Educational Research,* **30**: 28–35, September 1936.
26. FENTON, NORMAN. "An Objective Study of Student Honesty During Examinations," *School and Society,* **26**: 341–44, Sept. 10, 1927.

27. GLICK, M. N. *Effect of Practice on Intelligence Tests.* ("University of Illinois Bulletin," Vol. 23, No. 3, Sept. 21, 1925.)
28. GOLDSTEIN, H. "A Malingering Key for Mental Tests," *Psychological Bulletin,* **42**: 104–18, February 1945.
29. HARPER, BERTHA P., and DUNLAP, J. W. "Derivation and Application of a Unit Scoring System for the Strong Vocational Interest Blank for Women," *Psychometrika,* **7**: 289–95, December 1942.
30. HERBST, R. L. "How Accurately Do Teachers Score Achievement Tests," *Journal of Educational Research,* **20**: 140–47, 1930.
31. HERKELMANN, LEO EMIL. "A Study of the Use of the Separate Answer Sheets with Achievement Tests at the 4–6 Grade Level." Unpublished Master's thesis, State University of Iowa, August 1938. iii + 49 pp.
32. HEVNER, KATE. "A Method of Correcting for Guessing in True-False Tests and Empirical Evidence in Support of It," *Journal of Social Psychology,* **3**: 359–62, 1932.
33. HOFF, A. G. "Study of the Honesty and Accuracy Found in Pupil Checking of Examination Papers," *Journal of Educational Research,* **34**: 127–29, October 1940.
34. KELLEY, T. L. "Cumulative Significance of a Number of Independent Experiments: Reply to A. E. Traxler and R. N. Hilkert," *School and Society,* **57**: 482–84, April 24, 1943.
35. ———. "Scoring of Alternative Responses with Reference to Some Criterion," *Journal of Educational Psychology,* **25**: 504–10, October 1934.
36. KIRLIN, WARNER. "Motivation as a Factor in Achievement Test Performance." Unpublished Master's thesis, State University of Iowa, August 1938. 23 pp.
37. KOGAN, L., and GEHLMANN, F. "Validation of the Simplified Method for Scoring the Strong Vocational Interest Blank for Men," *Journal of Educational Psychology,* **33**: 317–19, 1942.
38. LESTER, HELENE, and TRAXLER, ARTHUR E. "Simplified Method for Scoring the Strong Vocational Interest Blank Applied to a Secondary-School Group," *Journal of Educational Psychology,* **33**: 628–31, November 1942.
39. LOPER, JAMES F. "A Study of the Use of the Separate Answer Sheet at the Third and Fifth Grade Levels." Unpublished Master's thesis, State University of Iowa, August 1939. iv + 47 pp.
40. LORGE, IRVING. "Tabulating and Test-Scoring Machines: Application of International Business Machines to Educational Research," *Review of Educational Research,* **12**: 550–57, December 1942.
41. *Manual of Instruction for the IBM Test Scoring Machine.* Endicott, N.Y.: Department of Education, International Business Machines Corp.
42. MANUEL, H. T., and KNIGHT, JAMES. "A Device to Facilitate the Scoring of Tests," *Journal of Educational Research,* **29**: 219–20, November 1935.
43. PINTNER, RUDOLF. "Accuracy in Scoring Group Intelligence Tests," *Journal of Educational Psychology,* **17**: 470–75, 1926.
44. POTTHOFF, E. F., and BARNETT, N. E. "Comparison of Marks Based Upon Weighted and Unweighted Items in New Type Examinations," *Journal of Educational Psychology,* **33**: 92–98, 1932.
45. PRICHARD, J. W. "Motor Performance as a Chance Factor in Test Scores," *Journal of Educational Research,* **37**: 181–92, November 1943.
46. RAUTH, J. "Scoring Objective Tests," *Catholic Educational Review,* **33**: 140–47, March 1935.
47. RUCH, G. M., and DEGRAFF, M. H. "Corrections for Chance and 'Guess' versus 'Do Not Guess' Instructions in Multiple-Response Tests," *Journal of Educational Psychology,* **17**: 368–75, September 1926.
48. RULON, PHILLIP J., and ARDEN, WESLEY. "A Scoring Technique for Tests Having Multiple Item Weightings," *Personnel Journal,* **9**: 235–41, October 1930.
49. SCATES, D. E. "Unit Costs in the Administration of a Standardized Test," *Educational Research Bulletin,* **16**: 38–45, February 1937.
50. SHEN, EUGENE. "Note on the Scoring of Matching Tests," *Journal of Educational Psychology,* **31**: 625–26, November 1940.

51. SODERQUIST, H. O. "New Method of Weighting Scores in a True-False Test," *Journal of Educational Research,* **30**: 290–92, December 1936.
52. STALNAKER, JOHN M. "Weighting Questions in the Essay Type Examination," *Journal of Educational Psychology,* **29**: 481–90, October 1938.
53. STENQUIST, J. L. "Experiments with Machine Scoring of Tests," *Baltimore Bulletin of Education,* **13**: 83–85, September 1935.
54. STRONG, EDWARD K., JR. "Weighted vs. Unit Scales," *Journal of Educational Psychology,* **36**: 193–216, April 1945.
55. TAYLOR, ERWIN K. "Some Suggestions for the Improvement of Machine-Scoring Methods," *Educational and Psychological Measurement,* **6**: 521–32, Winter 1946.
56. TINKER, M. A. "Speed, Power, and Level in the Revised Minnesota Paper Form Board Test," *Pedagogical Seminary,* **64**: 93–97, March 1944.
57. TOOPS, HERBERT A. "Code Numbers as a Means of Scoring Group-Administered Performance Test Products," *Journal of Applied Psychology,* **26**: 136–50, April 1942.
58. ———. "Directions for Administering and Scoring the Ohio State University Psychological Test, Form 20," *Ohio College Association Bulletin,* 1937, No. 107, pp. 2245–48.
59. ———. "A Proposal for 'A Standard Million' in Compiling Norms," Proceedings of a Conference of State Testing Leaders, October 28, 1939. (Mimeographed.) Washington: Committee on Measurement and Guidance, American Council on Education, 1939. Pp. 29–38.
60. ———. "The Use of the Hollerith Punch Machine in Scoring and Analyzing Tests of the Multiple Choice Form," *Ohio College Association Bulletin,* 1932, No. 67, pp. 752–57.
61. TRAXLER, ARTHUR E. "Accuracy of Machine Scoring of Answer Sheets Marked with Different Degrees of Excellence," *Proceedings of the Research Forum, Endicott, New York, August 26–30, 1946.* New York: International Business Machines Corp., 1947. Pp. 89–94.
62. ———. "Note on the Accuracy of Teachers' Scoring of Semi-Objective Tests," *Journal of Educational Research,* **37**: 212–13, November 1943.
63. ———. "A Procedure for Overprinting Answer Sheets for Hand Scoring Which Might be Adapted to Local Scoring," *Educational and Psychological Measurement,* **8**: 65–67, Spring 1948.
64. ———. "The Relation between Speed and Level of Literary Composition," in *1938 Achievement Testing Program in Independent Schools.* ("Educational Records Bureau Bulletin," No. 24.) New York: The Bureau, June 1938. Pp. 51–56.
65. ———. *Techniques of Guidance.* New York: Harper & Bros., 1945.
66. TRAXLER, ARTHUR E., and HILKERT, ROBERT N. "Effect of Type of Desk on Results of Machine-Scored Tests," *School and Society,* **56**: 277–79, Sept. 26, 1942.
67. TYLER, F. T., and CHALMERS, T. M. "Effect on Scores of Warning Junior High School Pupils of Coming Tests," *Journal of Educational Research,* **37**: 290–96, December 1943.
68. WALLEN, RICHARD, and RIEVESCHL, GEORGE, JR. "Improved Self-Marking Answer Sheet," *Journal of Educational Psychology,* **33**: 702–4, December 1942.
69. WEIDEMANN, C. C., and NEWENS, L. F. "The Effect of Directions Preceding True-False and Indeterminate-Statement Examinations Upon Distributions of Test Scores," *Journal of Educational Psychology,* **24**: 97–106, 1933.
70. WILKS, S. S. "Weighting Systems for Linear Functions of Correlated Variables When There Is No Dependent Variable," *Psychometrika,* **3**: 23–40, March 1938.
71. WOOD, BEN D. "Studies of Achievement Tests," *Journal of Educational Psychology,* **17**: 1–22, 125–39, 263–69; January, February, March, 1926.
72. WOOD, ELEANOR PERRY. "Improving the Validity of Collegiate Achievement Tests," *Journal of Educational Psychology,* **18**: 18–25, 1927.
73. ZUBIN, J. "The Chance Element in Matching Tests," *Journal of Educational Psychology,* **27**: 1–17, 1936.

11. Reproducing the Test

By GERALDINE SPAULDING
Educational Records Bureau

COLLABORATORS: Miriam M. Bryan, *Silver Burdett Company;* Herbert A. Toops, *Ohio State University;* Agatha Townsend, *Educational Records Bureau*

IN THE COURSE OF ITS EVOLUTION, A TEST SOONER OR LATER COMES to a stage where multiple copies are required. Quite a few copies are often needed for purposes of review in the course of construction; a fairly large number of booklets are required for experimental administration; and when all preparatory work has been completed, the test in its final form must be reproduced in quantity.

Kinds of Reproduction Process

Reproduction processes may be divided into three general classes: (1) duplication by office-type machines, such as the Mimeograph, Ditto, Multilith, and Multigraph machines; (2) printing by the photo-offset or planograph method; and (3) letterpress printing. Of these three, the first is usually the only one that is practical where test construction activities are limited to occasional and small-scale projects. The agency or individual responsible for the test ordinarily has access to a duplicating machine, and modest quantities of test materials can be produced by the regular office personnel with a minimum expenditure of time, effort, and money. Where more extensive programs are carried on, however, the scope of the work frequently justifies the use of the other two methods, in which the material is sent out to a commercial establishment for printing. The choice of the appropriate reproduction method will differ according to the requirements of the particular job, and each of the methods may be found most suitable for one or another project, or at one or another stage of a project. In addition, it sometimes happens that the various factors influencing the selection of the method for a given job do not all favor the same method, so that the choice is a matter of weighing and balancing. The following descriptions of the different methods are designed to furnish a general basis for making this choice, and therefore deal chiefly with the characteristics that have a direct bearing

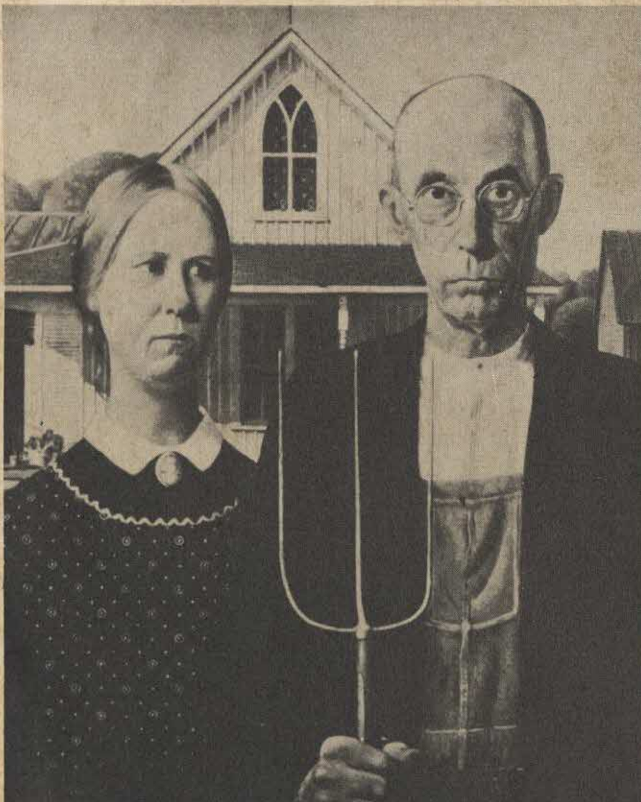
on the selection of the appropriate method in the light of particular job requirements. More detailed technical information can readily be obtained as need arises from representatives of printing companies and suppliers of duplicating machines.

Office-Type Duplicating Machines

When copies are to be produced on a duplicating machine, each page is typed on a special stencil, plate, or master copy. In the resulting copies, the reproduced material is of exactly the same size as in the original stencil or master sheet. The usual page size is standard typewriter paper size, 8½″ × 11″, or legal size, 8½″ × 13″. Limitations on the amount and arrangement of material on each page are the same as for typewritten copies. Drawings are made directly on the stencil, or are traced onto it, so that there is a practical limit to the amount and fineness of detail that can be included. When the test form contains more than one sheet (two pages, if sheets are printed on both sides), the separate sheets must be assembled and fastened together after duplicating. The number of copies that can be made from a single original typing varies according to the quality of the original typing and the age, condition, and make of the machine. While the multilith and the multigraph are well adapted to runs of several thousand copies, the necessity for the use of separate sheets and the consequent inconvenience in handling stapled booklets make it inadvisable to use these machines for the production of large quantities of test materials except when an entire test form requires only a single sheet. In general, a few hundred is generally the maximum number that it is worth while to produce by the use of office-type duplicating machines.

Photo-Offset Printing

Photo-offset printing yields exact reproductions of original copy. The original or master copy is photographed and transferred to large plates (a number of pages to each plate), from which the copies are printed. Original copy may consist of typewritten text, drawings, photographs, proofs or printed copies of material previously set in type, or any combination of such materials. Since a photographic process is used in transferring original copy to the plates, anything that can be photographed can be used in the original copy. Typed or printed materials, line drawings, and any other materials in simple black and white are reproduced directly. Photographs or other shaded illustrations can be reproduced by means of screening, which breaks areas in different shades of gray into



71. This painting represents which of the following tendencies among recent American painters?
    1) To follow the ideas of European painters
    2) To take their inspiration from American conditions of life
    3) To use painting as a means of bringing about some reform
    4) To use painting to express abstract ideas

FIG. 20.—Facsimile of original test item, illustrating photo-offset reproduction of a photograph (screened) and of typewritten material (pica type), reduced. (From 1941 Iowa Every-Pupil Test in United States History.)

patterns of dots in black and white. Color printing is possible but more expensive than black and white.

Offset printing has the advantage of permitting changes in size from the original, since copy can be either reduced or enlarged in the photographing process. Various parts of the same page can be reduced or enlarged separately so that, for example, illustrations which are not originally of an appropriate size can be made larger or smaller before incorporation into the page.

Typewritten copy is generally used as original copy for the text. For the sake of legibility, it is desirable to have the copy typed on sheets larger than the size desired for the final booklet pages, using a typewriter with type at least as large as the pica size (10 characters to the inch). When this copy is reduced in size in the photographing process, any minor irregularities in the outlines of the typed letters are also reduced, and the final printed page has a sharper, more clean-cut appearance than an unreduced reproduction. Figure 20, which is a facsimile of the original test item, illustrates a reduction of pica type. Appearance of the printed copies can also be improved by the use of special typewriters in making the originals. Standard or electric typewriters may be obtained with special type faces, some of which closely resemble letterpress type; and machines like the Vari-Typer are provided with interchangeable sets of type of different sizes and styles.

Printing is done on large sheets of paper from the plates carrying four, eight, or sixteen page units. Booklets are then made by folding the large sheets, wire-stitching through the center fold, and trimming the outside edges.

Letterpress Printing

Letterpress printing is the familiar process used in printing most books, magazines, etc. The printer sets the material in type, in accordance with the instructions furnished regarding choice of type and arrangement. Separate blocks are made for illustrations, and are inserted at the appropriate places when the pages are made up. The material sent to the printer for setting in type is ordinary typewritten copy. This copy may contain minor corrections and revisions made by hand, but the copy must be clear enough to be easily legible. It must give the exact text wanted if the expense of changes in the type is to be avoided. Indications of the size and style of type to be used, spacing, length of line, indentation, and other matters of format are included in the instructions to the printer. Drawings and illustrations of any kind can be reproduced,

in the same or a different size. An extra charge is made for illustrations; and the charges for composition (typesetting) are higher for tabular and other special materials than for straight text.

The printer furnishes proofs, which must be carefully checked. The specified corrections in the type are made by the printer, and second and third sets of proofs supplied for checking when required. When final approval has been given, the test goes to press. Booklets are made up by folding, stitching, and trimming, in the same way as when offset printing is used. It is possible to have the type kept standing or to have plates made, so that additional printings can be ordered at later dates. If errors are discovered after the first printing has been made, corrections can be made before subsequent printings.

Comparison of Methods

The following comparison of the three methods will summarize their respective advantages and disadvantages in the production of test booklets.

The office-type duplicating machine is the most practical and convenient method of producing straightforward material in relatively small quantities —that is, quantities ranging from eight or ten to several hundred copies. It is particularly useful when very little time is available for obtaining copies, since the process itself is not time consuming, and the work involved can often be performed entirely within the agency responsible for the test or in a closely affiliated office. The finished product, however, is less attractive in appearance, less convenient to handle, and usually less legible than printed booklets. Both offset and letterpress printing are superior to the duplicating machine for reproduction of material involving complicated diagrams, illustrations that are not easily traced or copied, or material which presents special difficulties in the design of the page.

Letterpress printing yields the most attractive and legible finished product. The printed impression is sharper and cleaner than offset reproduction of ordinary typewritten copy, even when the latter is reduced in size in reproduction. In addition, the more varied resources furnished by the styles and sizes of type commonly available provide greater flexibility in the design of the page and in setting off from one another the various elements that need differentiation.

However, time and cost considerations frequently favor the use of offset printing rather than letterpress. The two printing methods differ little with respect to the amounts of time required of the test agency's personnel, even when the preparation of original copy for offset printing is included; but a good deal more time is required for the actual printing

by letterpress than by photo-offset. The amount of time required for either method varies greatly from time to time and from company to company. In general, the offset printing of an eight-page booklet rarely requires more than a week or ten days, and in an emergency it is often possible to get 48-hour service. For letterpress printing, two weeks is about the minimum for such a booklet, and it is sometimes necessary to allow as much as four to six weeks.

Relative costs also vary a good deal. Offset printing is definitely cheaper for small quantities, and letterpress printing is cheaper for large quantities. The dividing line will of course depend on the particular price schedules in effect; but it will usually fall somewhere between 5,000 and 10,000 copies. The farther from the dividing line, the greater the differential: offset printing is very much cheaper than letterpress for 200 copies; and letterpress is very much cheaper than offset for 25,000 copies.

The decision about the method to be used for any given job should take into account the relevant considerations of time, cost, complexity of materials, and facilities available.

General Design of Printed Test Booklets

When the method of reproduction has been decided on, the next problem is the general design of the booklet. It is easy to dispose of this matter if a duplicating machine is to be used. The number of pages is unimportant, since separate sheets are assembled and fastened together; and there is little choice in page size and type size. Legal size paper ($8\frac{1}{2}'' \times 13''$) will take more material per page, and may thus permit the use of fewer pages than $8\frac{1}{2}'' \times 11''$ paper; but the longer sheet has the disadvantage of being somewhat more difficult to handle in filing and packing, since many standard containers and file drawers are designed for the $8\frac{1}{2}'' \times 11''$ size. The choice of type size is usually limited to elite or pica typewriter type, unless special typewriters are available.

However, when booklets are to be printed, there are various factors to be considered in planning the general design that are not involved in the case of duplicated copies.

Page Size

Printed test booklets may be produced in many different sizes. The most economical use of paper limits page size to certain conventional dimensions, dictated by the standard sizes of the large paper sheets used by printers. The most common size of page, $8\frac{1}{2}'' \times 11''$, is probably the best size for most purposes. Somewhat smaller sizes, such as $6'' \times 9''$

or 6″ × 10″, are occasionally desirable, especially for tests made up of single-line or other very short items; but the $8\frac{1}{2}$″ × 11″ page is more or less standard, since it provides ample space for pleasing and efficient arrangement of nearly all types of test materials, and still is not so large as to make an unwieldy booklet.

Number of Pages

Booklets made by folding and stitching have an even number of leaves or sheets, and therefore, since the leaves are usually printed on both sides, the number of pages is a multiple of four. It is possible to make up booklets with an odd number of leaves by stitching or gluing a single sheet in with the folded set, or by using parallel folding without stitching; but any saving is usually canceled by the increased cost caused by the departure from standard procedure. If the number of pages actually required for the test material (including cover page) is not an exact multiple of four, it is almost always best to make the number of pages equal the next higher multiple of four. The extra space can sometimes be advantageously used in more generous spacing of test materials on the page; or it can be distributed as blank pages at strategic spots. The outside back cover (last page) and the inside front cover (page 2) are good places for blank pages when the extra space is not needed for special purposes within the booklet.

Use of Separate Cover Page

A separate cover page is not essential if the directions to the examinee are short, and if strict timing is not important. The title of the test, spaces for the examinee's name and other information (if separate answer sheets are not used), and instructions for taking the test may be placed on the upper part of the first page, and the items may then begin on the lower part of that page. However, there is no particular advantage in this arrangement unless such use of the first page for items makes it possible to bring the total number of pages required down to a lower multiple of four. Suppose, for example, that the items will exactly fill seven pages. If the lower half of the first page is used for items, the only result is that the last page will have blank space at the end, where it serves no purpose. In such a case, it would be better to begin the items on page 2, reserving the cover page entirely for test title, information blanks, directions, sample items, etc. If, however, the space needed amounts to seven and one-half pages, the use of a separate cover page would make the last items fall on page 9. This would necessitate the use of a twelve-page booklet instead of an eight-page booklet, and involve increased expense

for printing. The use of half of the cover page for items would keep such a test down to eight pages.

This use of a part of the first page for items is inadvisable for tests where strict timing is important. The use of a separate cover page is obviously helpful in enforcing a rule that examinees are not to look at the items until the signal to begin is given.

Use of Blank Pages

Similar considerations apply to a test containing more than one part, when strict timing of the separate parts is desired. If a booklet can be arranged so that a part other than the last ends on a right-hand page, the beginning of the next part is not visible until the page is turned. A blank page inserted at the end of a part can be used to force the beginning of the next part over to a left-hand page if such a page arrangement would not otherwise occur. Even when a part does end on a right-hand page, a blank page can be used to advantage (if there are pages to spare in the booklet) by leaving blank the left-hand page following the end of the part and beginning the next part on the right-hand page. Where parts are several pages long, the beginning of a part is somewhat easier to locate on a right-hand page than on a left-hand page.

These paging arrangements are not essential if strictly uniform timing is not especially important. In such a case, the best procedure is to determine first the approximate number of pages required for the items and other essential printed material. If this number is an exact multiple of four, or very slightly less, an arrangement without blank pages or separate cover will keep the printing expense to a minimum. If the number of pages required is not an exact multiple of four, the best use of the extra space will depend on the particular test. Even where strict timing is not important, appearance is somewhat improved by beginning a part on a new page. In some tests, blank pages or parts of pages can be assigned to serve as scratch paper. In tests containing large groups of items referring to the same illustration, passage, or the like, a judicious distribution of blank space will help in arranging the material in such a way that it will not be necessary for the examinee to turn a page in referring back to the data on which the items are based.

Allowance for Directions, Samples, and Score Boxes

In estimating the number of pages required and planning the paging of the booklet, one should not overlook such material as general and specific instructions, statement of time allotments, sample items or practice exercises, and spaces for recording scores. In a test consisting of only

She noticed that there was a button missing from his jacket at once; she always noticed anything amiss with Ernie, because her pride—and she had plenty—was centered upon her boy. But, apart from that, Mrs. Poston was orderly in all her ways. She liked to have everything inside her cottage, from the shining row of neatly labelled tins above the mantle-shelf to the marker in the Bible by her bedside, arranged precisely in its proper place. Persnickety, she called herself—a neat, thin, white-haired, quiet-voiced little woman, the last person in the world to have a murderer for a son.

1. Mrs. Poston's habits of living were
   1 fanatical.
   2 careless.
   3 easygoing.
   4 fastidious. ....1( )

2. Mrs. Poston admitted that she was
   1 stern.
   2 fussy.
   3 peculiar.
   4 overly proud. ....2( )

Fig. 25.—Reading passage printed with lines too long for easy legibility

a single type of item, where a single set of directions and one or two sample items are sufficient, the directions and samples can well be placed on the cover. Where different directions, and possibly sample items, are required for different parts, they are placed at the beginning of the part. Where spaces for the scores are provided, they should be located so that the recording of the scores will involve as little turning of pages as possible.

Page Layout

Use of Columns

In planning the layout of individual pages, the first consideration is the length of line to be used for verbal material. Lines of text may run the

```
31.  One of the functions
     of white cells in hu-
     man blood is to
     31-1  destroy invad-
           ing germs.
     31-2  break down ac-
           id wastes into
           harmless com-
           pounds.
     31-3  liberate ener-
           gy for the mus-
           cle cells.
     31-4  carry oxygen.
     31-5  excrete excess
           moisture from
           the body.
```

Fig. 33.—Use of very short lines breaks up the text excessively

full width of the page, or the page may be divided into two or even three columns. The best arrangement is that which gives the optimum combination of legibility and economy of space, taking into account the page size, type size, and nature of test materials.

Legibility, insofar as it is affected by length of line, is primarily a matter of avoiding extremes, with respect both to actual length in inches and to the relationship between the length and the size of type (that is, the number of characters per line). Very long lines are difficult to read, particularly in such solid connected material as occurs in reading passages (see Figure 23) or in multiple-choice items having lengthy stems and

responses. When the eye has too far to travel from the end of one line back to the beginning of the next line, it is easier to lose one's place, and reread or skip lines, than when the lines are of moderate length. On the other hand, the use of very narrow columns is best limited to items having short elements, such as multiple-choice vocabulary items. If very short lines are used for running text, the division of a great many words at the ends of lines may seriously interfere with legibility (see Figure 22).

Economy in the utilization of space is chiefly a matter of selecting a page arrangement that makes due allowance for the foregoing considerations of legibility without leaving too much space unused. Where lines running the full width of the page are used for items having elements that, individually, do not take up most of the line, a good deal of space is wasted (see Figure 23). Such waste can be minimized without sacrifice of legibility by the use of a two-column page, provided the number of characters per line is not too small (see Figure 24). This condition is more often fulfilled in printed copies than in duplicated copies, since the duplicated characters are the full size of the typewriter type. Material for duplicating-machine copies should not be arranged in two columns if the small number of characters per column-line forces an excessive breaking-up of the text. In printed tests (either letterpress or offset), the greater sharpness of the type permits the use of somewhat smaller characters, so that the resulting larger number of characters per column-line is often just right for both legibility and economy.

Placing of Items on Page

Whatever page arrangement is adopted, each item should be given, complete, within a single column or page. It is particularly important that the examinee should not have to turn a page in the midst of any one problem. Items should never be broken at the bottom of a page or column and continued on the next page or column unless some other exceptionally important factor makes such breaking unavoidable. If a test is made up of items of different lengths, and if keeping the items in their exact order is not essential, it is often possible to avoid wasting too much space in following this principle by making minor changes in the order of the items. If an item is too long to fit into the remaining space in the column or page, it may be possible to exchange it with a later item that is short enough to fit. Or if the remaining space is insufficient for even the shortest item, the preceding item may be exchanged for a longer item that will take up all or most of the remaining space.

On a two-column page, the columns can be made approximately the same length by similar shifts. Where such shifts are impossible, or fail to make

1. You are living in a town near a river. In early summer there is a week of heavy rain. The stream rises, overflows its banks, and floods the town. Which of the following factors has probably contributed most to this disaster?

   A. Stream pollution
   B. The drying up of springs
   C. Widespread deforestation
   D. Accumulation of too much topsoil
   E. Fall of the ground water level

2. A single-celled organism is found to have cytoplasm, a nucleus, chloroplasts, vacuoles, and a cell membrane. Which of these indicates that the organism is a plant rather than an animal?

   A. The cytoplasm
   B. The chloroplasts
   C. The nucleus
   D. The cell membrane
   E. The vacuoles

FIG. 23.—Space wasted on lines containing choices

1. You are living in a town near a river. In early summer there is a week of heavy rain. The stream rises, overflows its banks, and floods the town. Which of the following factors has probably contributed most to this disaster?

   A. Stream pollution
   B. The drying up of springs
   C. Widespread deforestation
   D. Accumulation of too much topsoil
   E. Fall of the ground water level

2. A single-celled organism is found to have cytoplasm, a nucleus, chloroplasts, vacuoles, and a cell membrane. Which of these indicates that the organism is a plant rather than an animal?

   A. The cytoplasm
   B. The chloroplasts
   C. The nucleus
   D. The cell membrane
   E. The vacuoles

FIG. 24.—Double-column page permits more compact arrangement of same material

the columns exactly equal, the extra space in the shorter column may be distributed between the items so as to equalize the columns.

Care should be taken to keep such shifts in position within appropriate limits, as determined by technical or logical considerations governing the order of items.

Any reference material (passages of text, tables, graphs, etc.) should be on the same page as the items referring to it, or at least on a facing page. In dealing with such problems in the arrangement of pages, it should be remembered that even-numbered pages fall on the left and odd-numbered pages on the right of the two-page spread of an open booklet.

Provision for Response

When the examinee's response is to be indicated directly in the booklet (that is, not on a separate answer sheet), it is necessary to make specific provision for the entering of the response. In some tests (particularly those for younger children), the correct choice itself is underlined, encircled, or checked. However, as such responses are harder to score than responses written in a specified place, most tests provide spaces in which the choice number or letter is to be written. These spaces are printed in a straight column, to facilitate scoring. Parentheses or short lines are most often used to indicate the place where the response is to be written. This answer space is best placed at the right-hand margin of the column or page, since it is most convenient for right-handed scorers to indicate correct and incorrect answers with a mark in the margin immediately to the right of the answer. Placing the answer space at the right is somewhat better for the examinee as well, since the location of answer spaces at the left may require him to cover part of the item with his hand while writing his answer.

Answer spaces are sometimes placed at the left in spite of these considerations, presumably because the left-hand margin is always straight and answer spaces so placed will always be close to the edge of the text of the item. However, leaders (. . . . . .) can be used to connect the end of the item with the answer space; and the item number is sometimes repeated with the answer space. Either or both of these devices will make clear which answer space belongs to each item.

Use of wider margins at the *right* of each page (instead of the conventional use of equal margins on the *outside* edges of pages, regardless of whether the outside edge is at the right or the left) adds to convenience in hand-scoring, since it provides more space for marking the right and wrong answers.

Identification of Items

All items, of whatever type, should be numbered. Informal tests are sometimes reproduced without item numbers, since the test constructor feels no particular need for the numbers at the moment; but some means of identifying individual items is nearly always found desirable afterwards, and it is much simpler and more dependable to number the items as the copy is prepared than to supply identification later. Sometimes both numbers and letters are used, to indicate groups and individual items respectively. However, serial numbering of individual scorable units seems the simplest and most direct method of identification; groups can be identified in other ways—by marking off with lines or asterisks the separation between groups, by including at the head of the group some such note as "Items 6-9 apply to this diagram," or both. In tests having more than one part, each part may start with item 1, or the items may be numbered consecutively throughout the test. The latter plan is usually necessary if standard answer sheets are used, and has the advantage of furnishing positive identification of each item with a single number. For separately timed parts, however, separate numbering of each part provides the examinee with a better idea of his progress relative to the time allowed.

Page Numbers

Page numbers may be placed at the top or bottom, either centered or at the outside edge of the page. (The outside is at the left of even-numbered pages, at the right of odd-numbered pages.) On some tests designed for hand-scoring, the page number is placed in the upper *right* corner on both even- and odd-numbered pages, directly above the column of spaces for response, in order to facilitate matching the appropriate key strip in scoring.

Arrangement of the Individual Multiple-Choice Item

Closely related to arrangement of the page, of course, is the arrangement of item elements—reference matter (if any) such as reading passages, diagrams, tables, etc.; question, statement, problem, or in general "stimulus"; and suggested answers from which a choice is to be made.

Strung-out Arrangement

The arrangement of multiple-choice items shown in Figure 25 was formerly used extensively and is still used a good deal in informal tests made for local and limited use.

The only advantage of this arrangement is that it permits the use of

practically all the space on the page with a minimum of waste, since at most only a part of one line per item is left unused. Among the disadvantages is the fact that the lines of text are likely to be too long for easy legibility. A more serious disadvantage is that, in a long item, the individual choices are not well differentiated spatially. This makes it difficult

> 1. The electrical energy that can be obtained from a dry cell is (1) stored in the form of electrical energy, (2) stored in the cell in the form of chemical energy, (3) stored in the cell in the form of mechanical energy, (4) stored in the cell in the form of kinetic energy of molecules, (5) created within the cell...1( )

FIG. 25.—Item is difficult to read in strung-out form

for the examinee to consider each choice as a whole, and to compare and contrast the choices, especially when two or more choices have some words in common. In addition, the association of each label (letter or number) with the choice it identifies is not clear-cut and unmistakable when the label is buried within the text.

For these reasons, it is best to have choices strung out in this fashion only for items short enough to fit into a single line, with plenty of space between the choices (see Figure 26).

> 7. Which one of the following gases changes color on exposure to air?
> (1) NO (2) N_2 (3) Cl_2 (4) Br_2 (5) Ne......7( )
>
> 35. sheen A. spite B. luster C. organ D. turf E. plumage....35( )

FIG. 26.—Types of items for which horizontal arrangement is suitable

Arrangement in a Uniform Spatial Pattern

For longer items, stem and responses should be arranged in some spatial pattern that will clearly differentiate the choices from the stem and from each other. There are several such patterns; the selection of the appropriate pattern depends chiefly on the length and nature of the item materials (see Figure 27). In tests having items of varying nature and length, it is ordinarily advisable to select the one arrangement that is best suited to the majority of the items and use that same arrangement throughout the test or part, rather than to vary the arrangement from item to item. In most cases, there is more advantage in presenting a uniform

35. Which of the following reactions in the blast furnace produces carbon monoxide?

1) The reaction of the ore with the hot gases.
2) The decomposition of limestone.
3) The burning of coke.
4) The reaction between lime and the impurities.

36. If pure iron oxide is used as the raw material, which of the following need not be used in the blast furnace?

1) Limestone.
2) Coke.
3) Carbon monoxide.
4) Air.

Students entering a university are $\underset{24}{\underline{\text{liable}}}$ to be $\underset{25}{\underline{\text{amazed}}}$ by their new freedom and responsibilities and to waste a great deal of time before settling down to work. $\underset{26}{\underline{\text{Later they will regret it.}}}$ Since no one among all the people they meet $\underset{27}{\underline{\text{are}}}$ willing to tuck them $\underset{28}{\underline{\text{obligingly in}}}$

24. 1. liable 2. due 3. likely 4. prone

25. 1. amazed
    2. bewildered
    3. discomfited
    4. unadjusted

26. 1. Later they will regret it.
    2. They will regret it later.
    3. They will later regret it.
    4. OMIT

27. 1. are 2. is 3. will be 4. have been

28. 1. obligingly in
    2. in obligingly
    3. in
    4. away

1. The number **51** in the problem **17×3=51** is called

A. the quotient
B. the dividend
C. the product
D. the sum
E. the difference
F. none of these answers

49. In the formula $S=\dfrac{c}{c-x}$, c is a positive constant. As the value of x increases from zero and approaches c, the value of S

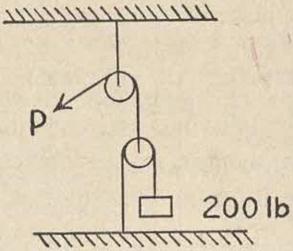
A. must increase
B. may remain constant
C. must decrease
D. may increase or decrease, depending on the value of c
E. becomes negative

FIG. 27.—Illustration of various ways of arranging multiple-choice items

spatial pattern for the item elements than in adapting the spatial pattern separately to each individual item.

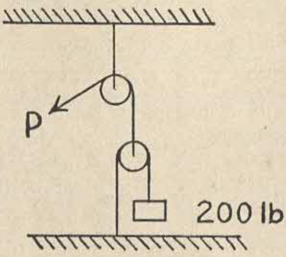
Where there is a diagram or drawing to which reference is made in the stem, the reference illustration is sometimes placed above and sometimes placed below the stem (see Figures 28 and 29). In general it seems more logical to put the illustration first, followed by the stem or question, in order to avoid separating the choices from the stem (especially when the stem is an incomplete statement).

43. If the efficiency of the pulley system shown is 30 per cent, the force required at *P* to lift the 200-lb weight is about



A. 111 lb
B. 167 lb
C. 333 lb
D. 1333 lb

FIG. 28.—Diagram interrupts verbal sequence from stem to choices



43. If the efficiency of the pulley system shown is 30 per cent, the force required at *P* to lift the 200-lb weight is about
A. 111 lb
B. 167 lb
C. 333 lb
D. 1333 lb

FIG. 29.—Stem and choices kept together

43. If the efficiency of the pulley system shown is 30 per cent, the force required at *P* to lift the 200-lb weight is about

A. 111 lb
B. 167 lb
C. 333 lb
D. 667 lb
E. 1333 lb

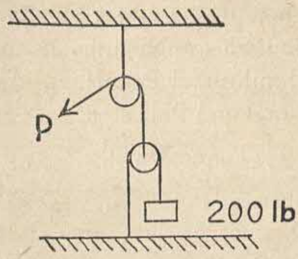


FIG. 30.—Compact arrangement, useful when economy of space is especially important

DIFFERENTIATION OF THE ITEM ELEMENTS

Where the choices are rather long, differentiation between item elements is helped by setting the numbers or letters identifying the items and choices out to the left of the body of text. In addition, stem and choices can be distinguished by the use of a little extra space between stem and first choice, and, if there is room to spare, between one choice and the next. Such extra space should, however, be less than the space between separate items, in order to preserve the effect of the total item as a unit. (See Figure 31.)

When typewritten copy is to be reproduced, the effects produced by variable spacing may be achieved, within limits, by the use of a typewriter having a special line-spacer that adds half-line spacing to the conventional single and double spacing.

Different amounts of indentation may also be used to differentiate item elements.

When tests are printed by letterpress, bold face type, italics, or different type styles may be used to set off the stem from the choices, or reference material from the item proper. However, it is well to avoid the use of boldface or italic type in large or numerous blocks of material, since such type is harder to read than the ordinary roman type, and since its effectiveness for purposes of emphasis is lost by excessive use. Similar effects may be obtained in typewritten copy for offset reproduction by the use of the Vari-Typer, or by using different typewriters with varying sizes or styles of type for different portions of the copy. In ordinary typewritten copy the means of differentiation are limited to the use of underlining, double underlining, and solid capitals.

In letterpress printing, boldface or other distinctive type faces may be

1. The legendary founders of Rome were
   1 Castor and Pollux.
   2 Damon and Pythias.
   3 Prometheus and Epimetheus.
   4 Romulus and Remus.
   5 Baucis and Philemon. ....1( )

2. The river on which the city of Rome is located is the
   1 Tiber.
   2 Po.
   3 Rhone.
   4 Rhine.
   5 Danube.................................2( )

3. Caesar's partners in the first triumvirate were
   1 Pompey and Crassus.
   2 Calpurnius and Clodius.
   3 Lepidus and Antonius.
   4 Cato and Marius.
   5 Brutus and Octavian.........................3( )

FIG. 31.—Item elements differentiated by use of varying amounts of space between lines

used to emphasize item numbers, and to differentiate the identification letters or numbers of choices. In typewritten copy, identifying letters or numbers may be followed by a period, enclosed in parentheses, or separated from the text of the choice by a single parenthesis or a dash (see Figure 32).

Labeling of Choices

There are several ways of labeling the suggested answers. Most commonly used are Arabic numerals and capital letters; lower-case letters are occasionally employed; Roman numerals are rarely used, and should be avoided. If responses are to be written by the examinee, either in the booklet or on a separate sheet, Arabic numerals are somewhat more likely to be legible than capital letters, and considerably more likely to be legible than small letters or Roman numerals. If responses are made by marking labeled boxes or spaces on a separate answer sheet, legibility of the examinee's response is not a factor. If numbers are used rather than letters, it may be a trifle easier for the examinee to avoid error in selecting the appropriate space on a separate answer sheet, since the number *4,* for

example, is more closely associated with the fourth space than the letter *D*.

A consideration in the labeling of choices that applies regardless of the method by which the examinee indicates responses is the avoidance of confusion between the label identifying the choice and the substance of the choice. If suggested answers to many of the items include Arabic numerals, as in arithmetic tests or other numerical materials, letters will

```
61.  Which one of the following states
     grows the most cotton?
     1.  Texas
     2.  Arizona
     3.  New Mexico
     4.  California
     5.  Florida. . . . . . . . . . . . 61(   )

61.  Which one of the following states
     grows the most cotton?
     (1)  Texas
     (2)  Arizona
     (3)  New Mexico
     (4)  California
     (5)  Florida . . . . . . . . . . . 61(   )

61.  Which one of the following states
     grows the most cotton?
     1)  Texas
     2)  Arizona
     3)  New Mexico
     4)  California
     5)  Florida. . . . . . . . . . . . 61(   )
```

FIG. 32.—Illustrations of ways of setting off choice numbers

be more distinctive as labels for the choices. If single capital letters appear in many of the answers, as in those involving chemical symbols, numbers will be more clearly differentiated as labels. Small letters are generally to be avoided, since capitals stand out better; but they remain a possibility for labeling in special cases.

ORDER OF CHOICES

It is more important to use the same system of labeling choices throughout any one test or part than it is to attempt to avoid all duplication between symbols used for labels and symbols used within the choices. Also, the use of standard answer sheets often dictates the way the responses are

labeled. In cases where duplication between symbols cannot be avoided, confusion can be minimized if the choices are so ordered that the choice label agrees with the choice itself. Numerical choices are preferably listed in order of magnitude, but it is generally considered better to list them out of order if that is necessary for placing, for example, a numerical answer "4" as choice No. 4, rather than as No. 2 or No. 5 (see Figures 33 and 34).

5. If a man spent \$.75 for socks and one pair cost \$.15, how many pairs did he buy?
   1. $\frac{1}{5}$
   2. 4
   3. 5
   4. 15
   5. 50

FIG. 33.—Answers arranged in order of magnitude, but likely to cause confusion between the answer proper and the identification number.

5. If a man spent \$.75 for socks and one pair cost \$.15, how many pairs did he buy?
   1. 50
   2. 15
   3. $\frac{1}{5}$
   4. 4
   5. 5

FIG. 34.—A better arrangement, avoiding confusion

Arrangement of Individual Items Other than Multiple-Choice

True-false and completion items are simpler to arrange than multiple-choice items, since they consist only of statements and simple provisions for response. The chief factors in determining arrangement are legibility and economy of space. The section above on page layout (pages 425–29) covers these considerations. The illustrations in Figure 35 show several possible arrangements of true-false items.

29. If a radius of the circle is drawn to the mid-point of an arc, it is perpendicular to the chord of the arc..................... 29( )

32. Gregor J. Mendel was one of the first scientists to use lenses in the study of small organisms.................................... ——

( )41. Eugenics is the study that seeks to explain the nature of differences and similarities between parent and offspring.

25. The New Testament was originally written in Hebrew...........T F

53. The English Reform Bill of 1832 was notable chiefly for its reform of the suffrage...........................TRUE FALSE

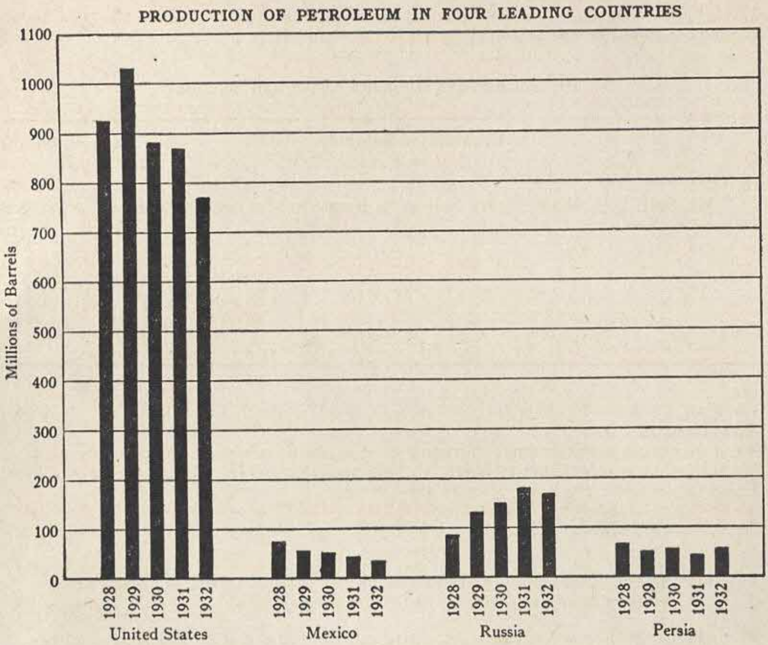
FIG. 35.—Illustrations of ways of arranging true-false items

Arrangement of Grouped Items

Groups Based on Common Reference Material

Grouped items of any kind present special problems. Where the basis of the grouping is simply the presentation of reference material that is to be used for a series of items, the problem is chiefly a matter of page layout. Where the group consists of material that will fit easily on one page, several arrangements are possible. If the reference material (reading passage, chart, map, etc.) and the items are of such proportions that they will fit within the same margins, the entire group may simply be presented serially, either in a single column or in two columns, with the items following directly after the reference material. If the reference material is wider than the items, it may be centered in the upper portion of the page, with the items arranged in two or three columns below. A long, narrow diagram may be placed at the left, with the items in a single column at the right. Other arrangements may be worked out to accommodate the particular combination of shapes and sizes presented by the materials. The chief factors to keep in mind are ease of reference from the individual items to the common basic material, and symmetry or at least balance in arrangement (see Figures 36 and 37).

If the group is too large to fit on a single page, as when a large map or drawing is used or when the number of items in the group is large, the best solution is to put the reference material on a left-hand (even-numbered) page, with the items referring to it on the following right-hand page. Where there are more items than will go on a single page, the reference material and some of the items may be placed on the left-hand page, and the rest of the items on the right-hand page.



(*Items 7 through 14 refer to the graph above.*)

7. The country that led in the production of petroleum in 1932 was
  1 **the United States**
  2 **Mexico**
  3 **Russia**
  4 **Persia** . . . . 7( )

8. The country that showed the most consistent rise in production was
  1 **the United States**
  2 **Mexico**
  3 **Russia**
  4 **Persia** . . . . 8( )

9. The country that showed the most consistent decrease in production was
  1 **the United States**
  2 **Mexico**
  3 **Russia**
  4 **Persia** . . . . 9( )

10. The number of barrels produced in Persia in 1931 was about
  1 **40**
  2 **4 million**
  3 **40 million**
  4 **175** . . . . 10( )

11. The country that showed the least fluctuation in production was
  1 **the United States**
  2 **Mexico**
  3 **Russia**
  4 **Persia** . . . . 11( )

12. The number of barrels produced in 1928 by all four countries combined was about
  1 **1 million**
  2 **400 million**
  3 **1150 million**
  4 **1388 million** . . . . 12( )

13. Russia produced less than 100 million barrels in
  1 **1928**
  2 **1930**
  3 **1931**
  4 **1932** . . . . 13( )

14. The two countries that produced about the same amount in 1930 were
  1 **Russia and Mexico**
  2 **Russia and Persia**
  3 **Persia and Mexico**
  4 **Mexico and the United States** . . . 14( )

Go on to the next page.

FIG. 36.—Illustration of page arrangement with reference material. (Cooperative Test of Social Studies Abilities, Experimental Form Q.)

Items 117 through 125 refer to the following experiment In studying the effect of various salts upon the relative humidity above water solutions of the salts, the following results were obtained:

Percent Relative Humidity Above Salt Solutions

Salt Added per 100 g H_2O	Unsaturated Solutions at 68°F						Saturated Solution at 68°F	
Formula	Wt. Salt	% Hum.	Wt. Salt	% Hum.	Wt. Salt	% Hum.	Wt. Salt	% Hum.
$CaCl_2$	20 g	80%	40 g	61%	60 g	44%	75 g	32%
$MgCl_2$	20 g	72%	40 g	48%	—	—	55 g	32%
NaCl	20 g	89%	—	—	—	—	36 g	78%
$Ca(NO_3)_2$	20 g	98%	40 g	95%	60 g	90%	129 g	59%
$NaNO_3$	20 g	96%	40 g	93%	60 g	86%	88 g	76%
NH_4NO_3	20 g	94%	40 g	88%	60 g	83%	192 g	68%

Using **only** these data, mark the degree of correctness of items 117 through 125 as follows

1 The statement is **true**
2 The statement is **probably true**, additional data would be necessary for a final decision
3 The statement is **impossible to judge**, the experiment provides no evidence upon which to make a prediction of the results to be expected in this case
4 The statement is **probably false**; additional data would be necessary for a final decision
5 The statement is **false**.

117. The relative humidity above a saturated solution of **KCl** is lower than the relative humidity above a saturated solution of **NaCl**.

118. At **68°F** at a concentration of **40 g** salt per **100 g** of **H_2O**, **NH_4NO_3** produces a lower percent humidity than **$Ca(NO_3)_2$** does.

119. These data were collected to explain why saturated solutions absorb moisture from the air.

120. At **68°F** increasing the concentration of the above solutions of these salts always tends to decrease the relative humidity above the solutions.

121. At **68°F** the relative humidity above a solution containing **30 g $NaNO_3$** per **100 g H_2O** is about **94.5%**.

122. At **68°F** the relative humidity above a solution containing **17 g NH_4NO_3** per **100 g H_2O** is **99%**.

123. The relative humidity above a solution containing **80 g $CaCl_2$** per **100 g H_2O** is **30%**.

124. On a weight basis, the listed chlorides are more effective than the listed nitrates in reducing the relative humidity at 68°F.

125. The relative humidity above a salt solution at **68°F** is less than that above pure water at the same temperature because the ions from the dissolved salts take up some of the surface area of the solution.

Go on to the next section.

FIG. 37.—Illustration of page arrangement with reference material. (Cooperative General Chemistry Test for College Students.)

If there are several smaller groups, each requiring considerably less than a page, reference material and items may be placed serially, with lines or asterisks to separate groups, or the page may be divided by lines or spaces into horizontal bands, with each group arranged to best advantage within its band (see Figure 38).

Groups Based on a Common Set of Choices

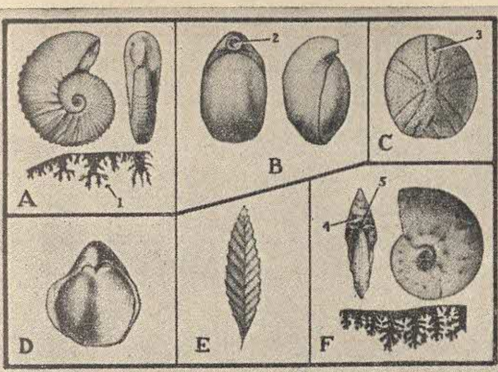
A different type of group is that in which the basis of grouping is the use of the same set of choices for a series of items. If the number of items does not differ greatly from the number of choices, the most economical as well as convenient arrangement is the placing of items and choices in two adjacent vertical columns. Items so arranged constitute what is usually meant by "matching items."

When the responses are to be written in the booklet, the three necessary columns (choices, items, and spaces for response) may be arranged in various ways: choices at the left and items at the right, followed by response spaces; or items at the left, with response spaces either before or after the items, and choices at the right. In the light of the scoring considerations discussed on page 372, the first mentioned is usually most convenient. If responses are to be placed on a separate answer sheet, it makes little difference which of the two possible arrangements of item column and choice column is used.

Another kind of group that is essentially a matching set involves a diagram or map with labeled parts that are to be matched with names, descriptions, or other associated phrases. Since the shape and size of the illustration are the controlling factors in the arrangement of such materials, the discussion on pages 437–41 of item groups based on reference material is more pertinent than the considerations applying to purely verbal matching items.

Still another variant of groups of items using the same set of choices is the combination of a small number of choices (usually from two to five) with a relatively large number of items. For such groups, the use of two adjacent vertical columns for choices and items is not economical, since a great deal of space would be wasted in the choice column. Such groups are best arranged by listing the common set of choices at the head of the group—perhaps integrated with the directions for the group—and arranging the items in sequence below (see Figure 39). The choices should be distinctly set off from one another and carefully identified with their labels. If the items in the group take up more than a page, the set of choices should be repeated at the head of each page. In many cases, the choices can be repeated in an abbreviated form which will serve as an adequate reminder of the marking code.

I



Referring to upper half of Figure I:
1. Brachiopod
2. Crinoid
3. Echinoid
4. Cephalopod
5. Pelecypod

10. A . . ( )
11. B . . ( )
12. C . . ( )

Referring to lower half of Figure I:
1. Cephalopod
2. Pelecypod
3. Angiosperm
4. Graptolite
5. Gastropod

13. D . . ( )
14. E . . ( )
15. F . . ( )

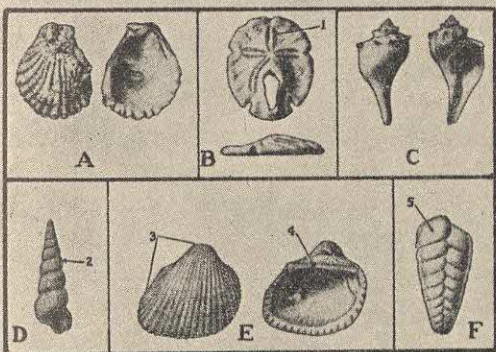
Place appropriate number from Figure I after each of the following:

16. Siphuncle. . . . . ( )

17. Suture . . . . ( )

18. Pedicle opening. . . . . ( )

II



Referring to upper half of Figure II:
1. Gastropod
2. Blastoid
3. Echinoid
4. Coral
5. Pelecypod

19. A . . ( )
20. B . . ( )
21. C . . ( )

Referring to lower half of Figure II:
1. Foraminifer
2. Gastropod
3. Pteropod
4. Brachiopod
5. Pelecypod

22. D . . ( )
23. E . . ( )
24. F . . ( )

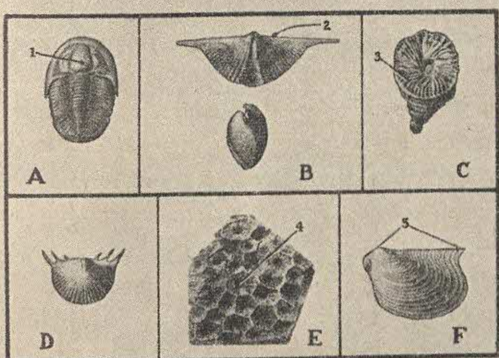
Place appropriate number from Figure II after each of the following:

25. Hinge line . . . . ( )

26. Whorl. . . . . ( )

27. Ambulacrum . . . . ( )

III



Referring to upper half of Figure III:
1. Sponge
2. Coral
3. Brachiopod
4. Bryozoan
5. Trilobite

28. A . . ( )
29. B . . ( )
30. C . . ( )

Referring to lower half of Figure III:
1. Coral
2. Bryozoan
3. Pelecypod
4. Gastropod
5. Brachiopod

31. D . . ( )
32. E . . ( )
33. F . . ( )

Place appropriate number from Figure III after each of the following:

34. Corallite. . . . . ( )

35. Glabella . . . . ( )

36. Septum . . . . ( )

Referring to Figures I, II, and III as a group:
1. Permian
2. Miocene
3. Devonian
4. Cretaceous
5. Cambrian

37. I . . . ( )
38. II. . . ( )
39. III . . ( )

Go on to the next page.

FIG. 38.—Illustration of page arrangement with reference material. (Cooperative Historical Geology Test.)

Items 88 through 93:

Four characteristics of fabrics are listed below.

Characteristics of Fabrics

1. Twill weave
2. Filling pile
3. Warp pile
4. Always yarn-dyed

Which of these characteristics is typical of each fabric listed below?

88. Gingham
89. Velveteen
90. Dress flannel
91. Tricotine
92. Transparent velvet
93. Gabardine

- - - - - - - - - - - - - - - - - -

Items 94 through 100 refer to the following list of articles of clothing which might be selected to produce a complementary color harmony with a navy blue suit.

Mark answer space 1 if the article should be selected.

Mark answer space 2 if the article should not be selected.

94. Dull orange felt hat
95. Warm beige blouse
96. Yellow gloves
97. Pink blouse
98. Chartreuse blouse
99. Powder blue blouse
100. Sapphire earrings

FIG. 39.—Illustration of arrangement of grouped items

Preparation of Copy for Photo-Offset Printing

Determination of Typing Area

In the preparation of typed copy that is to be reduced for photo-offset reproduction, the first step is the determination of the dimensions of the *actual typing area*. Margin width on the typed copy is immaterial and does not enter into these calculations, since the final margins depend only on the printed page size and the final printing area, and not on the amount of margin that the size of the typing paper happens to allow.

When the final page size, final margins, and amount of reduction have been decided (on the basis of the considerations discussed in the section on page layout, pages 425–29), the dimensions of the printing area are found by subtracting the margins from the final page dimensions. The reduction proportion is then stated in terms of the percentage of the typed copy dimensions that the printed copy dimensions will be. The desired printing area dimensions are divided by that percentage, the result giving the dimensions of the typing area.

As an example, suppose that the final printed page is to be 8½″ × 11″, with a ¾″ margin at the left and at the right, and a 1″ margin at the top and at the bottom, and that the typed copy is to be reduced 25 percent. Subtracting the margins from the page dimensions leaves a printing area of 7″ × 9″. With a 25 percent reduction of the typed copy, all printed-copy dimensions will be 75 percent of the corresponding typed-copy dimensions. Dividing 7″ and 9″ by .75, we get 9⅓″ and 12″ as the dimensions for the typing area.

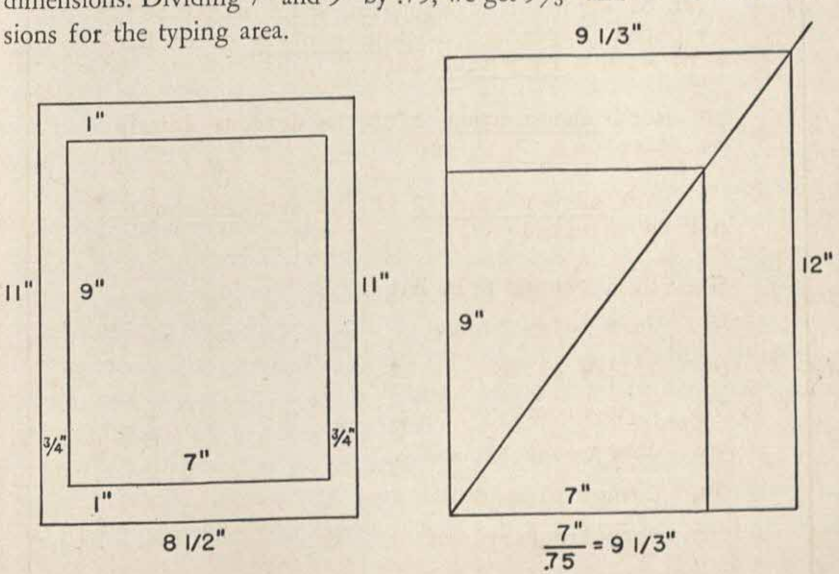


FIG. 40.—Diagram showing how to determine typing area for offset copy

Figure 40 shows this procedure graphically, using the same dimensions and amount of reduction as above.

Similar calculations may be carried out for any desired final page size, width of final margins, and amount of reduction.

A convenient size for the paper on which original copy is typed is 12″ × 15″. This size is ample to hold all the material in pica type that can be reproduced on an 8½″ × 11″ page with a reasonable amount of reduction.

When the typing area has been determined, it is helpful to have lines drawn or printed on the typing paper to indicate the space to be occupied by the typing. Offset printing companies often furnish sheets in standard dimensions with such lines printed, ready for use. Guide lines drawn on the paper should be made with a light blue pencil, to avoid having them show on the final copy; most blue marks will not photograph without the use of a special filter.

Provision for Illustrations

When final copy is being typed, space must be left for any drawings, photographs, or other illustrations. If drawings are to be made especially for the test, it is desirable to have them made in such a size that they can be reduced (in the same proportion) right along with the typed copy. Line drawings (that is, all black and white, with no shades of gray) that are to be reduced in the same proportion as the typed copy can be pasted down on the typed sheets with rubber cement (see pages 447–48); this will avoid extra charges for inserts, and will eliminate the danger of losing drawings or having them inserted in the wrong place.

Illustrations can, if desired, be reproduced with more or less reduction than the typed copy. They can also be enlarged, but this is rarely advisable, since enlargement cannot add detail that is not present in the original. In determining the appropriate size of illustration, one must consider both the loss of detail in reduction and the best size for combining with the text in an economical, pleasing, and clear arrangement.

When illustrations are *not* to be reduced the same amount as the text, the space to allow for the illustration on the typed copy has to be separately calculated, since there are three sets of dimensions involved: the size of the original illustration; the size of the final printed copy of the illustration; and the size of the space to be left blank in the typed copy. One should decide first what are the smallest final printed dimensions that will allow a sufficiently clear reproduction of necessary detail. The corresponding minimum dimensions of the space to be left on the typed copy can then

be found by the same method as that used in determining total typing area. These dimensions for the blank space are *minimum;* if a pleasing and economical page layout can be achieved with more room allotted to the illustration, the clarity of the reproduction will be increased.

It should be remembered that, whatever the amount of reduction or enlargement, both dimensions of any single piece of copy or blank space will be changed in the same proportion; that is, it is not possible to reduce the width by one percentage and the length by a different percentage.

Halftone illustrations (those showing gradations from black to white) and all illustrations that are not to be reduced in the same proportion as the typed text should *not* be pasted down, but should be sent to the printer in a separate envelope. Each illustration should be identified, with corresponding identification in blue pencil in the space marked off for the illustration on the typed copy.

Typing the Copy

Copies printed by offset are exact reproductions of the typewritten copy; it is therefore essential that the typed copy be not only free from errors, but clean, even, and sharp. Type must be kept clean, and the ribbon used must be one which makes a sharp, clear imprint. The ribbon must be black, and not too heavily inked. A silk ribbon usually gives good results. It is particularly important that an even touch be maintained in typing on an ordinary typewriter. A page that is darker in some places than in others will not photograph evenly, and light letters or parts of letters will show up as broken characters on the printed copies. One should watch particularly for light or missing "descenders" (the tails of letters like *g*, *y*, etc.) and blurred capitals. Light places in the typed copy can be touched up with a sharp-pointed, rather hard black pencil; but such touching up should be attempted only by someone with a very steady hand, taking care to get the pencil lines exactly in the track made by the typed character. Difficulties in producing copy of even darkness are minimized by the use of an electric typewriter or the Vari-Typer, since these machines provide mechanical control of the pressure.

The typewriter ribbon should not be changed in the midst of the typing. If a change cannot be avoided, it should be made at the end of a page, so that the difference in darkness will not be too noticeable.

In typing offset copy, it is easiest not to attempt to justify the lines (that is, make right-hand as well as left-hand margins even). The page does look more attractive when justified, but this involves a double job

of typing, since the extra spaces in short lines on the first typing must be distributed between words in the course of the second typing.

It is possible to combine typewritten and printed copy in a page to be printed by offset. If large, conspicuous page numbers, distinctive headings or labels, or other materials in regular type faces are desired, printed copies of the characters or words wanted may be pasted on the typed sheets. An old desk calendar is an excellent source for printed numbers and letters that can be cut out and pasted on the typed copy; and old tests, old magazines, and even newspapers are sometimes useful for special labels.

Layout for tests consisting partly or almost entirely of picture items is greatly facilitated by typing each item on a separate piece of paper and moving the items and pictures around on the various pages until a layout is achieved which meets most nearly the requirements for order of items, economy of space, and pleasing appearance.

Making Corrections

It is usually best not to make erasures, as the retyped characters are likely to be smudgy. On good quality paper, careful erasure of one or two characters will not affect adversely the appearance of the final copies, but the correction methods described in the next few paragraphs usually give better results.

For changes involving material up to several words, the best method of correction is the use of opaque white paint. A water-soluble paint such as show-card or poster paint (obtainable in small jars, ready for use) can be applied over the erroneous words with a small water-color brush and allowed to dry. When the paint is thoroughly dry, the correct letters can be typed on top of the paint. Care must be taken to have the white paint of the right consistency, and to apply a coat of the proper thickness. If the paint is too watery, or is spread out in too thin a layer, the original typing will show through. If the paint is too thick, or if too much is applied, it is likely to flake off when dry, particularly when typed upon. The paint is less likely to flake off if the retyping is done with a light touch; each letter may be struck two or three times if necessary to match the degree of darkness of the rest of the page.

For more extensive corrections, changes are best made by retyping the line or lines involved on a separate piece of paper and pasting the new copy over the old as a patch. In order to insure correct alignment, guide lines should be drawn with a blue pencil on both the patch and the base paper in precisely the same relative position. Adjustment of the patch so that the

horizontal and vertical lines are each continuous insures correct placement. It is a good idea to sharpen the blue pencil used for the guide lines on a sandpaper pad, since fine lines will give greater accuracy of placement. (See Figure 41.)

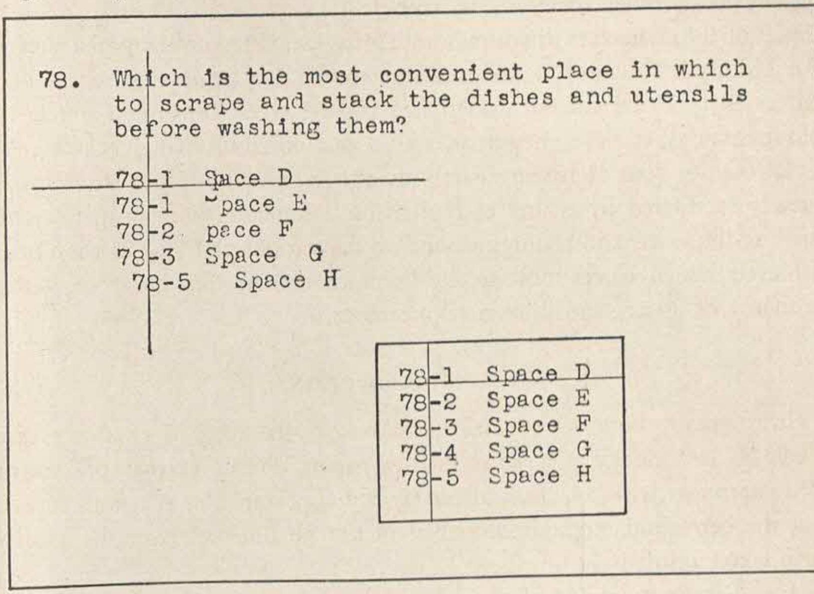


FIG. 41.—Illustration of use of guide lines to insure correct placement of patch on typed copy. Upper section represents portion of original page, with errors to be corrected; lower section represents patch with corrections. If guide lines in blue pencil are matched when applying patch, placement will be correct.

Do not use library paste or mucilage. These adhesives will wrinkle the paper, causing a distortion in the printed copy. Rubber cement, of the variety especially designed for use on paper, should always be used in pasting down correction patches and drawings, and in pasting together sections of pages.

To affix a patch permanently, one should apply the cement to both pieces of paper, taking care to spread the cement clear to the edge of the patch and to cover the corresponding area on the base paper. Cement applied only to the patch will hold it temporarily, so that the patch may be removed later if desired. In adjusting the placement of a large patch or a long, narrow patch, it is helpful to place between patch and base a piece of clean paper slightly shorter than the patch in one dimension and longer in the other. The ends of the patch can thus be adjusted first, with less danger that the patch will become permanently stuck before precise place-

ment is achieved. When the ends are properly aligned, the paper can be slipped out, and the rest of the patch pressed down.

Excess cement can be rubbed off with the fingers, after the patch is placed. In applying cement, care must be taken not to get it on typed material that is not to be covered with a patch, since some of the ink will come off with the cement, leaving a lighter area. Some people prefer to use a brush in applying rubber cement. Others like a small, flexible spatula. An artist's palette knife, in a small size, is a very useful instrument for this purpose. A wooden picnic spoon can also be used. A paper clip, partly unbent, is very handy for applying cement to narrow strips.

Minor errors detected by the typist in the course of the typing may be corrected at once, by erasing or by using white paint to blot out the error. More extensive corrections involving pasting should not be completed until *after* the proofreading of the test, although the typist may retype the copy for the patch and clip it to the page so that the new copy can be read in the regular process of proofreading.

Proofreading the Copy

All typewritten copy should be carefully proofread against the draft from which it was typed, preferably more than once. Corrections should be marked with a light blue pencil (which will not show in the reproduced copy). Smudges and marks that get on the copy by mistake can be covered with white paint. Page numbers and the numbers and letters identifying items and choices should be read *as a separate operation,* and they should be read *at least twice,* preferably by different proofreaders. Errors in identification numbers can invalidate whole blocks of responses, and nothing is easier to overlook than errors in such matter if it is not separately checked. Separate checks on spacing, margins, reference notes, parentheses or lines for responses, consistency of punctuation of items, correct answers to samples, and other such features are also desirable.

Particular care should be taken to make sure that patches have been proofread and found correct before they are pasted down. It is easier to type a new and perfect patch than it is to paste a patch down, find an error in it, and have to put a second (and even a third) patch on top of the original one.

All checking and proofreading should be done with great care, since there is no further opportunity for changes once the copy has gone to the printer.

Procedures for Letterpress Printing

Preparation of Copy

The preparation of copy for printing by letterpress is a much simpler job than preparation of copy for offset reproduction, since the material is to be set up in type instead of being photographed. A good legible typescript on $8\frac{1}{2}'' \times 11''$ paper, using one side only, is used for printer's copy. The typed copy should approximate, if possible, the desired arrangement and spacing of the final printed copy, in order to minimize the markings necessary to indicate to the printer the indentation, spacing, and so forth. Copy should be marked to indicate where boldface, italic, or solid capitals are to be used. Underlined words will be set in italic type; a wavy line under words indicates boldface type; double underlining indicates small capitals, and triple underlining indicates large capitals.

Copy should also be marked to show the size of type to be used. The unit used in indicating type sizes is the point. Figure 42 shows several sizes of type. Ten-point type is about the smallest that is easily legible for the body of the text.

This line is set in 8-point type.

This line is set in 10-point type.

This line is set in 12-point type.

This line is set in 14-point type.

FIG. 42.—Illustration of different sizes of type

Extra space between lines adds to legibility; text set with such extra space is said to be "leaded." Figure 43 shows text set solid (without leading) and Figure 44 shows text set with 2-point leading.

The normal man's dislike of his relatives lies, I believe, in the plain fact that every man sees in his relatives, and especially in his cousins, a series of grotesque caricatures of himself. They exhibit his qualities disconcertingly; they fill him with a disquieting feeling that this, perhaps, is the way he appears to the world. To admire his relatives wholeheartedly, a man must be lacking in the finer forms of self-respect.

FIG. 43.—Illustration of text set solid

> The normal man's dislike of his relatives lies, I believe, in the plain fact that every man sees in his relatives, and especially in his cousins, a series of grotesque caricatures of himself. They exhibit his qualities disconcertingly; they fill him with a disquieting feeling that this, perhaps, is the way he appears to the world. To admire his relatives wholeheartedly, a man must be lacking in the finer forms of self-respect.

Fig. 44.—Illustration of text set with 2-point leading

Headings and titles, such as names of parts, will stand out better if set in larger type than the body. Boldface, italic, or solid capitals can be used for emphasis, either in headings or in single words in the text that are to be stressed.

The typescript should be carefully proofread; every effort should be made to have it correct in all respects, since "author's alterations" involve needless expense.

Illustrations should not be attached to the copy, but should be placed in a separate envelope and enclosed in the package of copy for sending to the printer. Illustrations should be identified, and the place where they are to be inserted correspondingly identified. It is not necessary to leave on the typed copy the exact amount of space that the illustrations actually require; an inch or two of space can be left at the appropriate place, and a note added to identify the illustration that is to be inserted. It is, however, necessary to indicate the final size desired for the illustration. When printing is done by letterpress, separate "cuts" (electrotypes) are made for all drawings or photographs, and these can be made in any desired size. Here, as in the case of offset reproduction, one must consider two factors: How large *must* the picture be to show essential detail clearly? What size will combine best with the material set in type, for an economical and yet clear and pleasing arrangement of the page?

Paging the Material

It is usually best not to leave the paging of test materials to the printer, but to have the page divisions decided in the test construction office, since there are so many special considerations in paging a test that the printer cannot be expected to know about.

In the case of homogeneous materials with few or no illustrations, paging can best be done by working with the first set of proofs furnished by the

printer. These proofs, called galley proofs, are long strips of paper with single columns running continuously. An extra set of galleys may be requested for use in paging. If the test materials divide neatly into columns (if used) and pages without difficulty, it is sufficient to mark in the margins of the proofs the places where columns and pages end. Where several page layouts are possible, it is helpful to cut up the extra set of galley proofs and experiment with different arrangements. When a satisfactory layout has been achieved, the pieces of proof can be stapled or pasted to blank sheets of paper (with the item numbers corrected) and returned to the printer as an indication of the desired final page arrangement.

When dealing with material that offers particular difficulty in the arrangement of pages because of numerous illustrations of various sizes, or other lack of homogeneity in the test materials, it is helpful to have the printer's copy typed with the same average number of characters per line as the printed copy is to have. This can be done only when the printed copy is to use short lines, as in the case when there are to be two or more columns to the page, since a longer line will contain more characters than will fit on a standard typewriter page. The typed copy is made with a single column on each page. When the width of the typed copy (in terms of characters per line) is approximately the same as that of the printed page, it is much easier to judge the amount of space the printed text will require, and to make tentative decisions on the arrangement of the pages. Judgments about lengthwise space requirements can be made by counting the number of lines (in terms of single-spaced material) that the printed page will contain; the number of single-spaced lines on the typescript may be determined by measuring with a ruler, as typewritten lines run six to the inch. The chief reason for making such tentative decisions on page arrangement is that, in order to achieve a satisfactory arrangement, it may be necessary to make changes in the order of items, the size of illustrations, or even the general plan of the page. If such changes are required, it is much cheaper in money and time to make them on the typescript (by cutting and pasting or by changing instructions to the printer) than it is to make them after the test has been set in type and the cuts for the illustrations made.

Correction of Proofs

The galley proofs should be carefully proofread against the original copy, and all corrections clearly marked in the approved fashion. Some dictionaries contain a page showing proofreader's marks and the proper method of using them, and printers can usually furnish a similar sheet.

After the corrections have been made and the pages made up, the printer sends a set of page proofs for checking. These proofs should also be care-

fully proofread, with particular attention to the arrangement of material on the page and to material not provided for in the galleys, such as page numbers, directions about going on to the next page, etc.

Item and choice identification numbers or letters should be proofread as a separate operation; it is also well to check separately the spacing, provision for response, page numbers, punctuation, samples, etc.

If there are extensive changes, a second set of page proofs may be requested for a final check.

It is highly desirable to send a set of the galley and/or the page proofs to the original author of the test and to receive his approval before the final printing. If practicable, he should be asked to key the answers on his copy. This will provide a check on the key and insure his looking at the items carefully.

Printing of Accessory Materials

Much of the general discussion above on the reproduction of test booklets applies also to the printing of accessory materials. There are, however, a few additional points involved in the printing of answer sheets and scoring keys.

Machine-Scored Answer Sheets and Keys

Answer sheets for machine scoring are printed by the International Business Machines Corporation; such answer sheets are obtainable both in a number of standard forms, and in specially designed layouts. On answer sheets specially designed and printed for a particular test, the spaces to be used for recording responses must coincide with the standard locations, but any part of the sheet not needed for such recording of responses may be used for printing other material, such as special directions, norms, conversion tables, and even the items themselves if they are short.

On the IBM answer sheets, the answer spaces are arranged in ten blocks of 15 items each, with the scoring of each block separately controlled. In designing special answer sheets for tests on which separate part scores are necessary, it is desirable to arrange the answer spaces so that no block of anwer spaces contains items for more than one of the parts. If the answer spaces required for a part end in the middle of a block, the remaining spaces in the block should be omitted, and the next part started at the beginning of the next block. This may not always be possible. For example, a five-response multiple-choice test containing 150 items will just fit on one side of an answer sheet; if the test consists of several separately scored parts with the numbers of items in the parts *not* evenly divisible by 15, one must either use more than one side of an answer sheet, or else place items from

two parts in the same block. If the latter procedure is followed, it will usually be necessary to use more than one scoring stencil, and the scoring machine operator will need to make more than one insertion of the answer sheet into the machine in order to obtain all the part scores. The use of an answer sheet printed on both sides also requires an extra insertion, but the greater simplicity of the scoring keys required sometimes makes this procedure preferable. If technical difficulties arise in the designing of the answer sheet, it is well to seek advice from the IBM specialists.

For tests that do not require the use of all the answer spaces, this matter of part scoring is the first consideration in deciding which answer spaces to omit. A second consideration is that the most accurate scoring is obtained from the central portion of the answer sheet. For a very short test, therefore, it is a good idea to use the answer spaces nearest the center of the sheet.

Special scoring stencils for the IBM answer sheets may also be obtained from the International Business Machines Corporation. These stencils may be ordered with the name of the test and directions for scoring printed on them, and with any designated pattern of punched holes.

Standard IBM answer sheets may also be obtained which have a total of 30 answer spaces all in a single column along the right-hand margin. These are very useful for short tests and for the classroom teacher. The questions can be dittoed on the answer sheet itself. (They may not be mimeographed, since the carbon ink conducts electricity.)

Hand-Scored Answer Sheets and Keys

For scoring keys other than those used with IBM answer sheets, the best method for quantity production is probably photo-offset printing, since original copy can be prepared with the exact spacing desired and reproduced in the same size. Photo-offset is also probably most convenient and economical for separate answer sheets designed for hand-scoring (unless very large quantities are involved), since a draftsman can make up the original copy with the exact layout and spacing desired, and this original copy can be reproduced in facsimile.

Summary

This discussion of the procedures involved in reproducing test materials has emphasized the general features of the commonly used duplicating and printing processes, and the problems that most directly concern those who are responsible for test construction. Mention has been made of various general considerations in arranging for the reproduction of tests by office-type duplicators, photo-offset printing, and letterpress printing. In addition, more detailed attention has been given to general booklet design, page

layout, item arrangement, and the placing and reproduction of illustrations. Practical suggestions regarding the preparation of copy and proofreading have been made.

Whatever the method of reproduction used, the finished test booklet should present an appearance worthy of the care and attention that have gone into the construction of the test. Good format not only is pleasing to the eye, but also helps the test user to secure reliable results from the test. Poor page planning, with overlong lines of text in fine print, blurred or tiny illustrations, or confusing arrangement of materials can seriously impair the efficiency of test performance. If these hazards are removed and the format used to best advantage, the testing materials themselves are free to function.

Selected References

1. BENBOW, JOHN. *Manuscript and Proof.* New York: Oxford University Press, 1937.
2. *The Manuscript: A Guide for Its Preparation.* 3rd ed. New York: Wiley & Sons, Inc., 1941.
3. PATERSON, DONALD GILDERSLEEVE, and TINKER, M. A. *How to Make Type Readable.* New York: Harper, 1940.
4. SKILLIN, MARJORIE E., and GAY, R. M. *Words Into Type.* New York: Appleton, 1948.

For information about duplicating machines, consult the manufacturers' representatives. The following are perhaps the best known:

Mimeograph
Multigraph
Multilith
Davidson Dual Duplicator
Ditto

For information about letterpress and photo-offset printers, consult classified section of city telephone directory.

Many of the larger photo-offset printers issue excellent manuals on the preparation of copy for photo-offset printing.

12. Performance Tests of Educational Achievement

By DAVID G. RYANS
University of California at Los Angeles

NORMAN FREDERIKSEN
Educational Testing Service

COLLABORATORS: George K. Bennett, *The Psychological Corporation;* Harold O. Gulliksen, *Princeton University;* G. F. Kuder, *Duke University;* Leo Smith, *Rochester Institute of Technology;* Joseph Tiffin, *Purdue University*

FROM THE STANDPOINT OF VALIDITY ONE OF THE MOST SERIOUS ERRORS committed in the field of human measurement has been that which assumes the high correlation of knowledge of facts and principles on the one hand and performance on the other. Nevertheless, examinations for admission to the bar, for medical practice, for teaching, and even tests of ability to cook and sew, are predominantly verbal tests of fact and principle in the respective fields. Relatively little attention has been paid to the testing of performance as such.

Tests of information or knowledge are not, of course, to be discredited. They have an important place in education and industry for purposes of identifying certain kinds of individual differences. These tests are economical to use since they can be administered to large groups of persons and since they may be quickly and accurately scored. However, while they may provide important information about an individual's school progress, his general information background, and his knowledge of facts and principles, they often tell only part of the story. Many situations to which an individual is required to respond are very complex, and effective behavior in those situations demands something in addition to the knowledge of facts and principles.

Knowing the recipe for preparing food (the prescribed ingredients, the proper amounts of each, when each is to be introduced into the mixture, and the conditions under which the preparation should take place) does not, as any novice knows, assure the success of the finished product. Similarly, knowing the names and locations of the clutch, brake, accelerator,

steering gear, starter, and gear shift, and knowing the chronology of the operation of these parts does not insure, without practice, that an individual can drive an automobile. Still again, extensive knowledge of vocabulary and rules of grammar do not, in themselves, assure a student of the ability to express himself and his ideas in literary endeavors. It is for these reasons that performance tests are sometimes important devices for assessing educational achievement.

What Is a Performance Test?

Use of the Term "Performance Test" in the Measurement of Aptitude and the Measurement of Achievement

The term "performance test" has been used in connection with both the measurement of aptitude and the measurement of achievement. An *aptitude test* is commonly thought of as a device for measuring the capacity or potentiality of an individual for a particular kind of behavior. In the measurement of aptitude, previous experience or training on the part of the individual is assumed either to be lacking or to be constant for all individuals comprising the population considered.

It is in the testing of aptitude rather than achievement that performance tests probably have been most frequently mentioned in the literature of psychology. Thus, performance tests have been used extensively for the determination of the general intellectual background or general level of ability of individuals who suffer from language deficiencies. In this sense *performance test* is more or less synonymous with "nonverbal test" and is used primarily to distinguish this type of measurement from that requiring ability to comprehend and respond verbally. Familiar examples of such performance tests of general intellectual ability are those requiring reproduction of patterns of tapping (the Knox Cube Test), those involving the fitting of blocks or forms of differing geometrical shapes into the proper depressions in a form board (the Seguin Form Board), and those involving the tracing of a maze (Porteus Maze Test). Such tests have been widely used in seeking to estimate the capacity for learning of deaf children, foreign-born persons who are unfamiliar with the language in use, and persons who have not had the benefit of instruction in reading and writing.

Other performance tests of aptitude of a somewhat different nature and purpose have been used in the testing of capacity for training with respect to specific skills. Tests of this type are not used primarily as a substitute for intelligence tests, but rather to supplement paper-and-pencil tests of verbal and quantitative abilities in determining the individual's neuromuscular co-

ordination, manual dexterity, spatial perception, etc. Data derived from such tests have been found to be useful in the prediction of ability to learn various mechanical skills and operations.

An achievement test, as contrasted with an aptitude test, presupposes training and is intended to provide a continuum upon which the relative proficiency of different individuals at a particular sort of acquired behavior may be judged.

Performance tests of achievement purport to provide objective means for estimating the proficiency with which a task is performed. The situations involved are functional to a high degree and are likely to be complex. Because of the complexity of the situation, administrative control is often difficult. In spite of such problems associated with the measurement of performance, however, performance tests of achievement have been successfully developed in certain areas of behavior. Perhaps the most familiar tests of this type are those used in determining the progress of students of typewriting. The objective determination of errors, and of units of work produced per unit of time, serves to reveal the accuracy and speed of typewriting performance. Other commonly known performance tests of achievement are those used by state motor vehicle departments in the licensing of automobile drivers. In such tests the operator is required to start, stop, and maneuver the automobile as prescribed by an examiner. Additional examples might be drawn from the fields of industrial arts education, music, and other areas.

It should be noted here that the measurement of aptitude and the measurement of achievement can never be considered entirely separately. Achievement presupposes aptitude, and it is usually impossible to measure aptitude except in terms of previous experience of the individual, even though such experience may not be highly specific. Although this inherent relationship is recognized, the present chapter will, nevertheless, make a practical distinction between the two (aptitude and achievement) for purposes of discussion, and henceforth will consider the topic of performance tests exclusively as they are used in the measurement of achievement.

The Kinds of Performance Tests of Achievement

Performance tests are not at all new in the measurement of achievement. Athletic competition has a long history dating back at least to the Greek games and races 800 years before Christ. Athletic games, meets, and tournaments employ the principles of performance testing to reveal differences in skill of trained individuals or teams. Similarly, musical performance and literary and artistic endeavors have been judged and rated for many years.

Although the techniques often may have lacked refinement, and although the control of conditions of administration has not always been adequate, the procedures employed actually were those of performance testing.

More recently, performance tests have been used in industry in determining the proficiency of an employee, or prospective employee, at a particular skill. In these industrial situations, performance tests have been devised for the measurement of ability to "do the job" or to produce some industrial product.

One of the first applications of performance testing in education came with the adaptation of certain psychophysical methods for the judgment of quality of handwriting. The usefulness of performance tests of achievement is obvious in vocational curriculums involving the teaching of such skills as typewriting, machine operations, and food preparation.

In general, performance tests may be divided into three major types: (1) *recognition tests;* (2) *tests involving simulated conditions;* and (3) *work sample tests.*

Recognition tests

The recognition type of performance test, as the label implies, attempts to measure the individual's ability to recognize essential characteristics of a performance or product of performance, or to identify objects such as geological or botanical specimens.

A musical selection is played on an instrument and the examinee is required to indicate errors or deficiencies in execution or interpretation. A series of splices of electrical wires is presented and the examinee is required to determine those that are correctly done and those that may be inadequate. The examinee is presented with a piece of mechanical equipment into which certain defects (breakage of parts, poor adjustments, etc.) have been introduced, and he is required to locate and identify the defect. All of these are examples of performance tests involving the recognition of the wrongness or rightness of equipment, a process, or a product.

In other performance tests of the recognition type, the examinee is asked to identify mechanical parts and their functions in a particular assembly, to choose the proper tool or equipment for a defined operation, or to judge the quality of specimens of material or work. In the natural sciences students may be required to identify certain geological specimens, or to describe trees from leaves, twigs, and bark. Similarly, students of art or literature may be required to judge and select superior and inferior artistic and literary productions.

Performance tests of the recognition type are relatively easy to prepare and are adaptable to a fairly wide range of situations. Although they meas-

ure important aspects of performance, they do not measure directly the individual's mastery of a skill, technique, or procedure.

Tests involving simulated conditions

Performance tests sometimes are designed to copy or simulate the real-life situations or operations that the test is devised to measure. Such tests seek to isolate and duplicate the essential activities of an operation or task. From the examinee's performance in this representative situation, judgment is made regarding his ability or skill in the real situation.

"Simulated conditions tests" sometimes have been referred to as miniature tests.[1] Such a test as used in industry, for example, involves apparatus that has been especially constructed for the purpose. Although the test situation is artificial in a sense, it has certain advantages of administration (mock equipment may be provided in multiple sets for testing purposes with greater economy and convenience than can the more expensive and often scarce machinery), and of safety (use of the specially constructed device minimizes dangers that might be involved in the operation of a production machine by a novice or unskilled worker who is being tested).

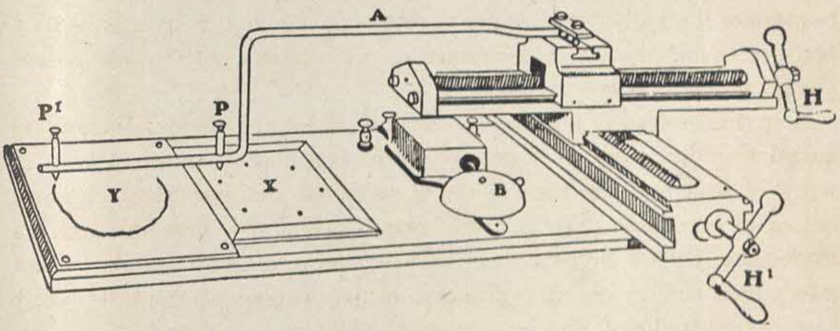


FIG. 45.—Wisconsin Miniature Test for Engine-Lathe Operation. (Reported by Patten [*20*, p. 451]; illustration reproduced from Viteles [*30*, p. 229].)

A number of tests involving simulated conditions have been devised. Figure 45 shows the Wisconsin Miniature Test for Engine-Lathe Operation reported by Patten (*20*). The test apparatus is so constructed that various aspects of the examinee's behavior may be recorded in a convenient manner.

A miniature punch press for measuring ability to operate an industrial punch press, such as described by Tiffin and Greenly (*27*), is shown in

[1] Viteles (*30*) makes a distinction between "miniature" tests and "simulated job" testing, defining the former as a test which reproduces the job in miniature and the latter as a test that simulates the job without reproducing it.

Figure 46. Figure 47 shows the Vigilance Test described by DeSilva and Channell (*8*) and used for measuring steering, braking reaction, and combined braking and steering in automobile drivers.

It should be noted that although the use of simulated conditions tests and miniature tests in the measurement of performance has certain advantages, the method must be used with caution and with as complete

FIG. 46.—Miniature punch press test. (Reproduced from Tiffin and Greenly [*27*, p. 451].)

knowledge as possible of the relationship between the results of the miniature test and the more complete performance with actual equipment and under actual conditions. In the training of gunners' mates for the Navy, for example, the situations involved in repairing a 20-mm. gun on a work table in a land-based training school may be quite different from those that apply when repairing the same gun on a pitching and rolling ship under battle conditions. During World War II the Army Air Forces found that a performance test for bombardiers consisting of bombing desert targets made up of concentric circles, or even a simulated factory or building, was

not satisfactory as a measure of an individual's accuracy in bombing a defended target in an actual air raid. And in schoolteaching, performance tests requiring a prospective teacher to conduct a strange class through a day's study may result in quite different estimates of the teacher's ability as compared with judgments and evidences based upon the continuous conduct of a class throughout a term or year.

FIG. 47.—Vigilance test used for measuring operations and reactions of automobile drivers. (From De Silva and Channell [8, p. 61].)

Obviously then, what may seem to be a valid sample of an isolated task or procedure—the miniature test—may not be a valid sample of the situation *in toto*. This may be due to a number of causes. The physical conditions of the simulated test may not actually reproduce those of the performance in question. Furthermore, it may not be possible to duplicate important psychological conditions of the examinee, such as those related to emotional components of the situation (which often affect performance to a considerable degree) in the simulated test.

Simulated conditions tests of performance, then, may have a useful place in achievement testing, but their limitations must be recognized and overcome if they are to serve the purposes intended.

Work sample tests

The work sample test of performance consists essentially of a "controlled" tryout under the actual conditions of the work situation. The

examinee is required, under normal conditions, to carry through the operations that the job demands.

The work sample test illustrates the "identical elements" test of chapter 5. This type of test is realistic; it has greater face validity than any other type of test or examination. When administered under standard conditions with standardized scoring procedures that have been carefully worked out, the work sample test may provide valid and reliable estimates of achievement for many kinds of behavior or performance.

In use, the work sample test may include the complete sequence of behavior or operation required by a given job or piece of work, or it may consist only of selected samples of job behavior. Obviously, for many types of activity the former is not economical of time and expense. In general, the more limited the sample of behavior that will predict the whole of that behavior, the greater the advantage so far as economy in test administration is concerned. Therefore, it is common practice to seek selected samples of performance that are sufficiently predictive of the behavior as a whole to insure adequate measurement of individual proficiency. The problem of sampling in the development of work sample tests will be discussed in greater detail later in this chapter (pages 467–69).

Work sample tests are of two principal kinds: (1) those in which a clear-cut distinction between the "rightness" or "wrongness" of the execution of a skill is possible, and which, therefore, are more or less automatic in scoring; and (2) those which must depend upon the judgment of observers for evaluation and assignment of a score or rank. Target shooting, foot races, mechanical assembly tests, and typewriting tests generally fall into the first category. They are capable of very objective scoring. In contrast, the proficiency demonstrated by an individual at automobile driving or violin playing, or the quality of performance reflected by some product of that performance such as a painting, a novel, a wire splice, or a chest of drawers, are measurable principally in terms of judgments made by presumably competent observers and with the aid of adapted psychophysical methods.

Illustrations of work samples are readily available. The classroom English theme is a work sample of the student's ability to think and express his ideas in verbal form. A pattern employed by the Blum Sewing Machine Test (*4*) is shown in Figure 48. It is designed so that the basic elements of sewing may be directly measured with a work sample. (In this test the examinee is required to sew on the line in situation A and between the lines in situation B. The line is zigzagged to resemble changes in direction required in actual sewing jobs.)

Again, when a student of music is required to play or sing a musical

selection for purposes of judgment and criticism, his performance is a work sample.

Among the best-known work sample performance tests are those which

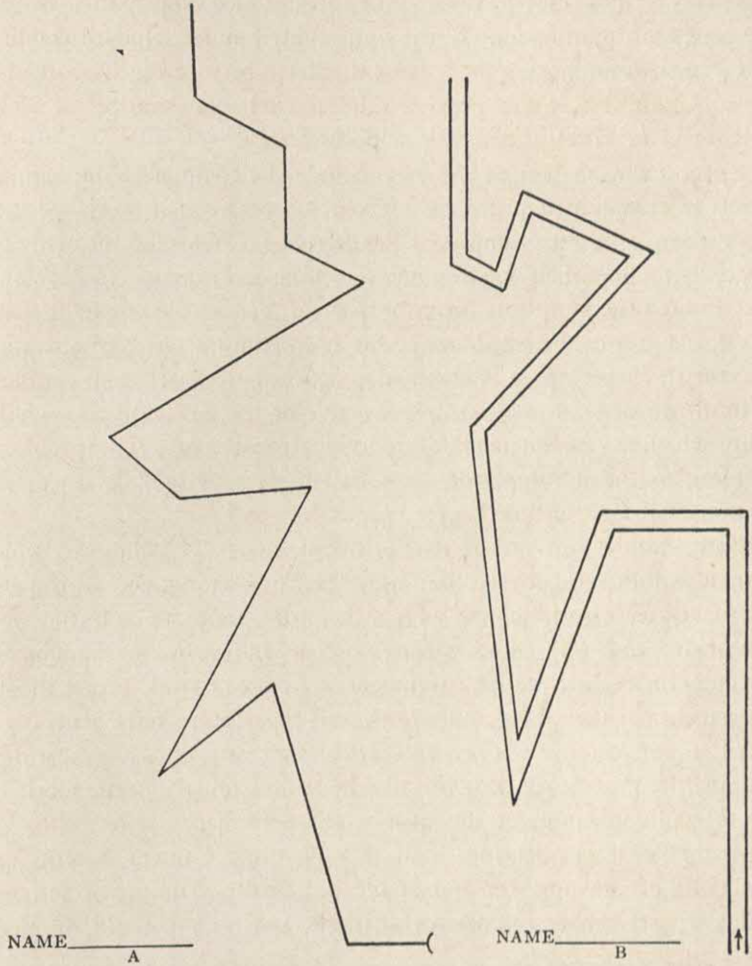


Fig. 48.—Blum Sewing Machine Test. (Reproduced from Blum [*4*, p. 36].) Reduced one-half.

require stenographers to take dictation and transcribe their shorthand notes, and those which require the typist to prepare a sample of copy for grading with respect to quality and speed of performance.

Uses of Performance Tests of Achievement

The immediate purpose of performance tests is the measurement of individual differences in proficiency of performance. But the ultimate purposes

or *applications of performance tests* are the principal interest of education and industry. In general, the uses of performance tests of achievement may be classified in four categories: (1) their use for the prediction of successful execution of skills; (2) their use in the diagnosis of deficiencies in performance; (3) their use as a teaching aid; and (4) their use for providing a criterion measure.

For the Prediction of the Successful Execution of Skills

The most obvious use of a performance test is in the determination of the relative proficiency of individuals with respect to a particular procedure or operation—the estimation of job effectiveness. Such estimates of proficiency have important applications in education, industry, and everyday life.

One such use is in the *selection* of personnel in industry. It is generally true that the selection of employees can be significantly improved if the results of an appropriate performance test, or tryout at the operation or job for which the candidate is applying, are employed. A carefully prepared performance test may approximate quite closely the task required in a job situation.

A second use of proficiency data yielded by performance tests is in educational and industrial classification and placement. Since students and employees vary greatly with respect to proficiency at a given skill or type of job behavior, it is important for purposes of economy of time and cost that individuals be "placed" as nearly as possible at the level of their achievement. In the school, and for industrial training, *further instruction should always begin at the level already attained by the individual.* In job placement, *both efficiency of operation and individual satisfaction are enhanced* by placement according to level of ability.

Still another use of determinations of proficiency yielded by performance tests is in the licensing of operators of mechanical equipment. Perhaps the best known of such procedures is the driving test for automobile operators.

For Diagnosis of Deficiencies in Performance

The use of performance tests for the diagnosis of deficiencies in performance also is important. Basically, diagnosis involves an analysis of aspects of behavior (of an individual's strength, of his arithmetic skill, of his industrial productivity, etc.) into observable units and the assessment of the quality of the behavior thus scrutinized. In the diagnosis of achievement the performance test provides standardized conditions for such analysis and for the determination of difficulties or deficiencies in behavior.

A number of years ago Freeman (9) called attention to the possibility of diagnosing faults in handwriting through the measurement of performance. The handwriting performance of the individual (writing arm too near body, thumb too stiff, index finger pressing too heavily, movement too slow, too much lateral movement, etc.) was observed and certain of the difficulties identified.

As a Teaching Aid

Although this use is one that perhaps is less commonly recognized, performance tests may contribute significantly to educational programs as *teaching aids*. They are of particular importance in that they provide motivation to the student through the direct revelation of success or failure in execution of the function toward which training is directed. And, of equal importance, they provide the instructor with a clean-cut means of evaluating his teaching—that is, revealing areas in which the students are, or are not, attaining desired objectives. Both experiment and experience have called attention to the effect of "knowledge of progress" upon learning. The performance test provides a direct and unequivocal indication of satisfactory or unsatisfactory performance and also stimulates the student to try to improve his skill.

During World War II, performance tests were found to be particularly effective in bringing about the improvement of instruction and learning. Despite the fact that skills required of service personnel were usually of a distinctly practical nature, instruction in the service schools was sometimes dominated by lectures and verbal descriptions of apparatus and its operations, or at best by demonstrations which were *assumed* to be useful in acquainting the trainee with the necessary techniques or understanding. Even when "learning by doing" was undertaken, its purpose was often defeated by the lack of individual supervision and attention required to avoid wrong learning. The introduction of performance tests in many of these schools focused attention upon familiarity with equipment and on the skills and understanding of procedures necessary for its proper maintenance, operation, and repair.

Regardless of what the teacher or course syllabus may describe as the course objectives, the efforts of the students will be directed mainly toward passing the examinations. The *course objectives,* therefore, *are likely to be achieved to the extent that they are reflected in the testing program.* Recognition of this fact, provided it results in examinations which truly reflect the objectives of the course, will go a long way toward improving learning. If a course or curriculum has as its aim the development of skills other than

those of a verbal nature, the use of performance tests is essential to effective learning.

Providing a Criterion Measure

When the behavior involved in a situation is broad enough and representative enough of the situation as a whole, the performance is itself the criterion behavior for that situation. Consequently, performance test data, particularly when they refer to work samples, provide a more satisfactory measure of criterion behavior than is usually available. Because performance tests serve as a measure of the criterion, they may be of use in several important ways.

Performance test data may provide, first of all, *a criterion for research.* Information yielded by performance tests makes possible the validation of other measures which, although of a more indirect nature, may be more convenient for use and more economical in administration. In many situations it is difficult and expensive to administer performance tests to large numbers of examinees. Such situations demand the construction of psychometric instruments that will yield measurements related to the criterion and which also will be practicable. In the construction of aptitude tests for various skills and operations, performance tests may provide the criterion against which the available second-order measures can be judged (*19*).

Again, performance test results may be used as a *criterion for advancement* or promotion of individuals in training and in service. The most usable index of present level of attainment, and, by inference, of the qualifications of the individual for successive levels of training or job activity, is provided by the performance test.

Finally, performance tests provide a *criterion for the general evaluation of a training program* in school or industry. Since the primary purpose of many educational activities is the progressive development of skill, the accomplishments of such a program can best be evaluated in terms of the proficiency in such skills attained by the students.

Development of the Performance Test

Many performance tests have a high degree of "face validity"; they appear to duplicate the job situation so closely that there seems to be no question that they are measuring the intended abilities and measuring them well. Face validity alone, however, is no guarantee of real validity or of high accuracy of measurement. An individual might, for example, drive an automobile over a standard driving course with satisfactory proficiency, but still experience considerable difficulty in the more confused emotional and

operational situation involved in driving his automobile along certain outlet routes at five o'clock in the afternoon in New York City. While the performance test, requiring operation of the vehicle over a standard course, may have sampled satisfactorily certain aspects of driving ability, the sampling may have been too narrow to be sufficiently valid. Again, a typist who has drilled intensively on a particular paragraph of copy, or a pianist who has practiced diligently on a particular selection for a recital, may satisfactorily "pass" a performance test involving skill at the particular copy or musical selection studied. Such a test probably would not be a valid or reliable test, however, of the one individual's typing skill or the other's musical accomplishment. Performance tests must be submitted to the same empirical checks of reliability and validity as are paper-and-pencil tests of aptitude and achievement, and with the same degree of rigor.

The Problem of Sampling

The validity of a performance test will depend to a large extent on the particular tasks which are chosen to be included in it to represent the more general abilities which the test is designed to measure. The choice of these tasks should be made in the light of thorough knowledge of the job as a whole, as exemplified by a job analysis, as well as upon various practical considerations.

In selecting sample tasks for inclusion in a performance test, limitations often are imposed by such factors as time and amount of equipment and personnel available. If it is desired to develop a performance test to measure ability of trainees in disassembling and assembling Diesel engines, where only four engines are available for fifty trainees, and where the total job might require several hours for a team of men working together, building a test which can be individually administered to all trainees within a reasonable amount of time and with the supervision of only two or three instructors would seem to be impossible. But it is in situations of this sort that performance tests are badly needed for such purposes as the evaluation of training, the motivation of trainees, and individual guidance and placement.

The problem may be solved by choosing parts of the total job to represent the task as a whole. In choosing jobs to be included in a test, the test constructor probably would not select them on a random basis or on the basis of the proportion of the total time devoted to certain jobs. Much of the task might be of a routine nature, such as assembling nuts and bolts to hold a housing in place. It probably would be more desirable to choose the salient parts of the task, those which are especially important because they are difficult or because they are crucial to the proper operation of the equip-

ment. One should be guided in this selection by a careful job analysis, by the opinions of experts, and by his own experience in learning to perform the task.

In the interests of economy of time, it would be desirable to include in each task a minimum of easy, routine operations. There is little to be gained by requiring an examinee in a test situation to remove and replace a whole series of nuts and bolts when one or two such operations might be sufficient to demonstrate knowledge of the correct procedure.

As many of the really salient tasks should be included in the performance test as is possible with the time and equipment available. If amount of equipment is a limiting factor, this condition may sometimes be overcome by breaking each piece of equipment up into subassemblies, so that different people can work on the same equipment simultaneously.

Various specific recommendations have been made for the selection of samples for performance test situations. Chapman (6), for example, makes several suggestions for choosing a work sample for a performance test in industry. He states that: the operation should be sufficiently exact to admit of accurate standardization and to enable objective judgments to be made; the task chosen should have face validity, and should be of a nature to command respect and establish the confidence of tradesmen; the materials, tools, and equipment should be reduced to the smallest practical quantity and should be capable of standardization so that all tests may be given under uniform conditions; the performance should not require an undue length of time; the performance should involve as little repetition of identical procedures as possible; and a preliminary tryout of a performance test should always be made using experts as the examinees to detect possible problems or difficulties. Adkins and Primoff (*1*) suggest further that the sampling of activities should be as wide as is practical; that each workpiece (test situation) should be used to test as many of the activities involved in the behavior as possible; and that the work sample should be designed for ease of measurement with available devices as well as with a view to its representativeness.

If the number of separate jobs which can be included in a test situation is small and represents only a small proportion of the possible jobs, then coaching in these specific tasks to the exclusion of the job as a whole becomes a possibility. Instructors who are interested in having a good test record might be guilty of such coaching, or information about the tests might be passed on from one generation of students to the next. The difficulties in this problem can be minimized by developing sets of alternative forms of a test which in the aggregate cover all the important features

of the total job. Then if information as to which form of the test will be given is withheld, students and instructors must be prepared for all possibilities.

On the other hand, it may be that there are only a few really important features in the total job, and that all of these can be included in the test. If this is true, then there is no objection to specific coaching, and it would in fact be desirable for trainees to know in advance exactly what comprises their test. One of the characteristics of a good performance test is that practice or "cramming" for the test results in improvement in ability to do a job. In contrast, cramming for an oral trade test probably would result in little or no improvement in job performance.

The ideal method of investigating the validity of a performance test is, as is true of any achievement test, to study its relationship to a suitable criterion measure. The difficulty, of course, is that a criterion that is more satisfactory than the performance test results themselves is usually lacking. The development of a criterion based on "on-the-job performance" is highly desirable as a check on the success with which good judgment has been employed in selecting appropriate tasks for inclusion in the test. It will also be of direct assistance for the development of satisfactory scoring standards.

Evaluating the Results of Performance Testing

The scoring of a performance test obviously will depend upon the kind of skill being measured. The relative importance of speed, accuracy, use of approved methods, or quality of the product must be weighed, and a scoring procedure must be developed which will adequately reflect the decisions reached. In filing a block of metal to meet certain specifications, for example, one may ask if it is more important that the job be done quickly, that the specifications be exactly met, that the file be held in a prescribed manner, or that the finish of the product be free of scratches. The relative importance of these and other characteristics of the performance must be decided, and each must be adequately represented and properly weighted on a score sheet. The decision as to the nature of the judgments to be made will ordinarily be based on the evidence from job analysis, expert judgment, and one's own experience in learning to perform the job. Final evaluation of the procedures developed should ideally be based on the relationship of the various test "items" to a satisfactory external criterion.

In many situations the choices are offered of measuring *performance in process* or of measuring the results of the performance, the *product*. Thus, in the hand-tool shop an individual's performance in filing a piece of metal

may be judged either from the way he holds the file, the kind of strokes he uses, etc., or it may be judged from the quality of the finished product. Similarly, in typewriting, the typist's position, touch, etc., may be observed and rated; or the typed copy may be graded for conformance to prescribed standards. Even in schoolteaching performance there is often a choice between observing and rating a teacher as she guides her pupils in learning, and judging the proficiency of her teaching from the product, the improvement in learning evidenced by the pupils (in relation to their ability).

Measurable characteristics of the performance in process.—What are the measurable characteristics of performance in process? A number of specialized scoring methods might be mentioned, but these resolve themselves generally into two major categories: (1) those relating to the estimation of quality of performance; and (2) those having to do with the speed, or rate, of performance.

In some situations one of these characteristics is definitely more important than the other. In many kinds of behavior they are both important and serve to complement each other. Thus, in the familiar example of typewriting it is highly desirable that the typist prepare copy *both* accurately and rapidly. A rapidly typed letter that is inaccurate will be of little use; on the other hand, since a principal purpose in using a typewriter is to increase efficiency, there is also a premium on speed.

If success on the job is dependent largely on the speed with which it is carried out, as in taking dictation or loading magazines on a 20-mm. gun, then obviously speed should figure strongly in the scoring of the test. If errors are important, then appropriate penalties should be subtracted or separate speed and error scores reported. If quality is the primary consideration, as in machining metal parts to close tolerances or in playing a musical instrument, then speed should be given little or no weight in the score.

It might be mentioned here that in certain types of training situations the use of speed scores has proved to be an excellent motivating device, even though improving speed was not considered a major objective of the teaching. Competition for low time-scores among gunner's mate trainees in Navy schools, for example, proved to be an excellent incentive for practice in the disassembling and assembling of guns.

In the measurement of speed or rate of production, the primary scoring concern is that accuracy and uniformity in counting and timing be attained. The use of automatic counting devices, such as the Veeder counter, is often possible. In other situations the counting process is facilitated by assembling the finished products in cases or racks that are constructed to hold a given

number of units. For timing purposes, chronoscopes or kymographs may be used, but the stop watch will be satisfactory for most situations where production rate is a major concern.

Time required to produce a single *unit of work* may be used as a measure of production rate, particularly when the time required to produce a sample is relatively long, but it is usually less satisfactory because of the greater unreliability of this method than counting the *number of units produced during a given time.* In either case an average time required to produce a unit of work may readily be computed. Timing separate units does, of course, have the advantage of making possible the computation of a measure of variability in production rate.

Although quality and speed of performance are the two major measurable characteristics of performance in process, it may be noted that performance is sometimes judged from such additional correlates of performance as learning time, accident rate, attitude toward job, etc. These will not be discussed here in greater detail.

Measurable characteristics of the product of performance.—In measuring the product of performance, the principal characteristics to be considered relate to the *quality* of the product. In measuring quality, certain standards (standards of appearance, usability, etc.) are adopted and the product is judged in relation to such standards.

In the case of a piece of metal turned on a lathe, for example, the product must conform to prescribed measurements or dimensions in order to serve the purpose intended. Somewhat similarly, a typing specimen will be judged in light of certain standards of spacing, position on the page, and freedom from error. Thus, conformance to prescribed dimensions, standards of general appearance, freedom from error, strength, and suitability for use may be used to judge the quality of a product and, by inference, the quality of the performance of an individual.

Often it is possible to objectify the measurement of quality of the product through the use of patterns, gauges, and scoring keys; the product may be classified, and scored, as it does or does not conform to the pattern or key. For some work samples such automatic scoring is less feasible, and it is necessary to resort to the judgments of qualified observers of the relative goodness or poorness of the product as it is compared with some agreed-upon standard or as it is compared with other similar products.

Should measurement of performance be directed at the process or the product?—While performance may often be measured either in process or in terms of the product, some situations are more limited and offer less choice. Occasionally the product of performance *is not* measurable apart

from the performance in process. Playing a musical instrument is an example of such a case. Driving a car (unless driving record over a period of time, and in terms of such recordable data as number of fender dents, number of accidents, or number of traffic violations, is considered) is another. In still other situations, analysis of the process is difficult and it is generally not possible to measure the performance in process except by judging the product of the performance. Examples of this sort are the writing of musical selections, the composition of literary products, etc.

In other instances, *both* the performance in process and the product of the performance may be measured by judging an individual's proficiency at a given sort of behavior. Certain advantages, and also certain disadvantages, are attached to each.

For example, it will be recalled that while the objective measurement of performance in process is possible in some situations, many times it is necessary to resort to subjective judgments which tend to be unreliable. Even when such ratings are refined and when they are made by trained and competent judges, there may be uncertainty as to whether the reported differences in performance are due to differences between the performances being judged or to the variability of the judges. For such reasons, it may be of doubtful value in some situations to attempt to measure the performance in process.

On the other hand, it *is* highly *desirable* to measure *actual operation*, or performance in process, in many cases. The final product of performance may appear to be of satisfactory quality, but the operational methods or procedures employed may have been unsatisfactory, even to the extent of creating a hazard to personnel or equipment. A taxi driver, for example, may successfully and speedily negotiate a route through traffic, but may at the same time constitute a problem for other drivers and pedestrians along the way. In such an instance, some attempt must be made to measure performance in process.

Again, as has already been noted, it is impossible in some situations to measure the product of performance apart from the process. Musical performance provides a case in point. Such performance may be recorded, it is true, but measurements of that performance must always relate to judgments made *during* the execution. Examples of a similar nature indicate the obvious need for measures of performance in process.

In general, measurement of the product of performance is likely to be somewhat more reliable than the measurement of the performance in process. Through the use of patterns, gauges, graded sample quality scales, etc., a relatively high degree of accuracy of measurement can be attained.

Another advantage of the measurement of the product of performance over the measurement of performance in process lies in the relatively more convenient administrative procedures of the former. Fewer proctors or administrative assistants are required for judging the product, and lesser demands are made upon them. Usually the product can be judged or graded at any time following completion of the performance.

However, an important consideration in measuring the product of performance has to do with the irreparability or irrevocability of errors or flaws that may have been introduced at some early stage of workmanship on the product. Such errors may be impossible to correct and may influence all later operations. While a major portion of the workmanship may have been of superior quality, the early mistake may detract seriously from the product as a whole.

Techniques for the measurement of performance in process.—In the measurement of the *quality* of performance in process, the methods available are largely subjective in nature and usually consist of ratings of behavior made by presumably competent judges.

Objective recording devices have been constructed in a few instances. For example, Barnes and Amrine (*3*) describe a method and apparatus for automatic scoring, in terms of time and errors, of the performance of driving a screw down into a tapped hole. Behavior of the subject of the workpiece is transmitted over an electric circuit and a record of performance produced on a kymograph. Similarly, Lindahl (*15*) developed a procedure for analyzing the foot movements of cutoff machine operators, obtaining an objective recording of individual patterns of performance on a moving paper tape. With a certain amount of ingenuity similar automatic devices employing the kymograph, Veeder counters, or other available recording instruments may be developed.

For the most part, however, variations of the psychophysical method of single stimuli, or absolute judgment, have been employed in the past, and remain a principal source of data, in scoring quality of performance in process. The performance is simply "rated," with reference to some specified continuum, by a qualified observer. In practice, the use of such methods ranges all the way from relatively superficial judgments of the "satisfactory-unsatisfactory" type to fairly detailed ratings which provide for the breaking-down of operations and the judgment of separate processes or operations entering into the performance. Perhaps the most common of these procedures is that which is based upon analysis of the performance into "steps" or operations and which makes use of a point-scale with certain values arbitrarily assigned for the successful completion of given

operations. A rough scale[2] of this sort, devised for judging ability to use a saw, is shown in Figure 49 (*21*).

In measuring efficiency of performance in terms of the *time required to complete* a phase or cycle of the operation, various timing devices may be used. In some instances very exact measurements requiring precision

To Saw to a Line with a Rip and Cross-Cut Saw

Tools and Materials: Sharp rip saw and cross-cut saw, bench, wood vice, and piece of wood.

Directions: Observe pupil as he works, and rate him on the following points

1. *Clamping stock:* 1 2 3 4 5 6 7 8 9 10
   Stock should be so held that it will not be loosened or cracked, and that its position will facilitate sawing.
2. *Starting cut:* 1 2 3 4 5 6 7 8 9 10
   With thumb at line, saw should be placed against the thumb. Saw should be pulled back slowly a few times to make a groove, then pushed forward.
3. *Holding saw:* 1 2 3 4 5 6 7 8 9 10
   Saw should be held firmly. For cross-cut saw, angle should be 45 degrees; for rip saw, 60 degrees.
4. *Stroke:* 1 2 3 4 5 6 7 8 9 10
   Stroke should be long and even, not too fast. Proper angle should be kept during sawing. Line should be followed.
5. *Ending cut:* 1 2 3 4 5 6 7 8 9 10
   The piece being cut off should be held with the free hand. Saw strokes should be slow and with little pressure so as to prevent breaking off the end.

Fig. 49.—A rough point-scale for judging ability to saw to a line with a rip and cross-cut saw. (From Proffitt, Ericson, and Newkirk [*21*].)

timing may be necessary. However, under most conditions of performance testing the use of a stop watch for timing will provide entirely satisfactory and useful score units. When the speed of performance of a group of examinees is to be tested, it may be desirable to utilize the services of a central timekeeper who, in addition to giving the starting signal for the test, will, at intervals, indicate the amount of elapsed time. This method

[2] The question may be raised whether or not such an arbitrary rating device properly may be called a "scale." Certainly the requirements of a scale as it is viewed technically (an established zero point and equality of intervals) are lacking. However, sufficiently high reliability is often obtained with such relatively crude instruments, and if such is the case, there is justification for their use in practical situations.

was used to advantage in the administration of performance tests in Navy gunnery schools during World War II.

Techniques for measuring the product of performance.—A number of devices have been used for measuring relatively accurately the *quality* of the product of performance insofar as such activities as typing, handwriting, woodworking, electrical wiring, mechanical assembly, etc., are concerned. Error scores have been used in typewriting, patterns and gauges in wood and metal work, and graded sample quality scales in handwriting, electrical work, metal work, etc.

In judging the quality of many products that must conform to prescribed dimensions, devices such as rulers, combination squares, Vernier scales,

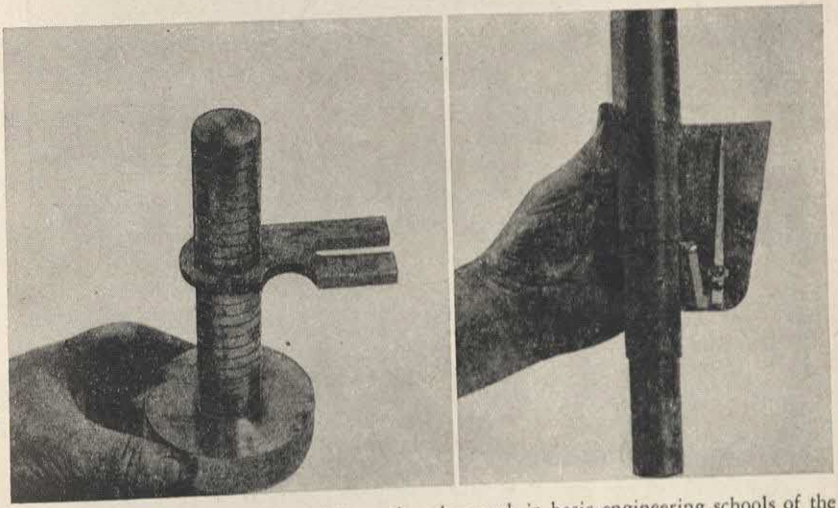


FIG. 50.—Two gauges employed for rating shopwork in basic engineering schools of the U.S. Navy. (From Stuit [*24*, Fig. 4-xv].)

and micrometer calipers may be employed. However, the use of such instruments, which require relatively precise reading, may be very unreliable (due to observer errors) particularly when a fairly large number of samples are being measured. Lawshe and Tiffin (*14*), for example, found the use of precision measuring instruments in industrial plants very inaccurate.

In order to facilitate and objectify the evaluation of a product, patterns and gauges often may be developed, permitting relatively automatic scoring. Such instruments may be devised to indicate clearly on a "pass-fail" basis whether or not certain standards are met by the product. Figure 50 shows two gauges employed for rating products of shopwork in the basic engineering schools of the United States Navy (*24*).

In Figure 51 is shown a relatively simple device revealing "wind" or unevenness of a flat surface (*19*). Figures 52 and 53 show a "dimension meter" and a "squareness machine" (*19*). These devices were constructed in developing criterion measures for tests of mechanical ability.

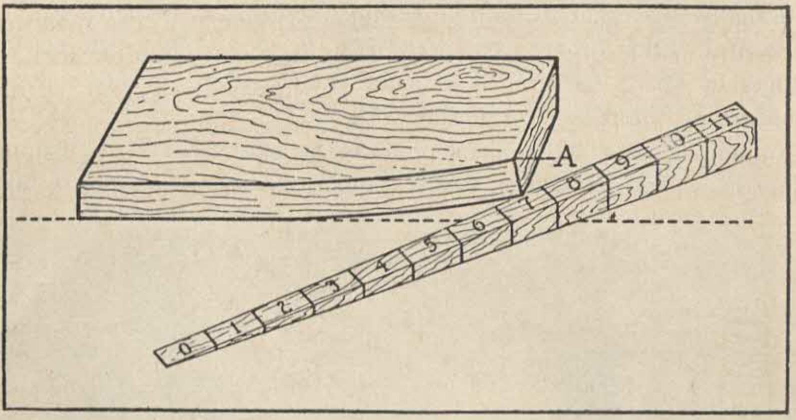


FIG. 51.—Simple device for revealing "wind" or unevenness of a flat surface. (From Paterson and Elliott [*19*, p. 190].)

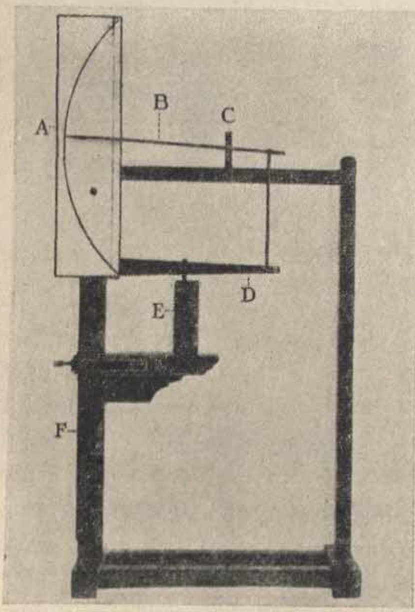


FIG. 52.—Dimension meter for testing mechanical ability. (From Paterson [*19*, p. 190].)

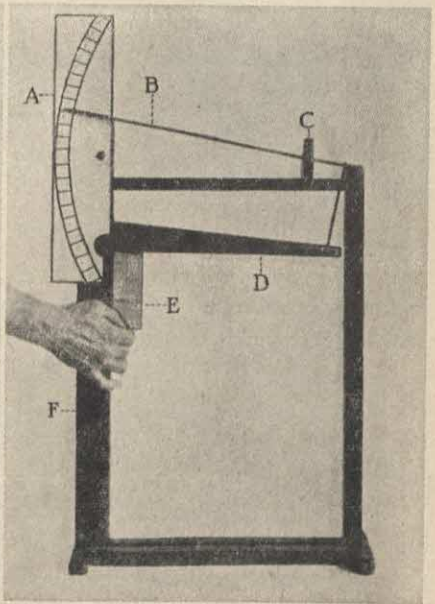


FIG. 53.—Squareness machine for testing mechanical ability. (From Paterson [*19*, p. 190].)

Mechanical and electrical devices also may be employed in determining the relative quality of the product. Thus, a soldered joint may be measured for its conductivity with special types of galvanometers. Similarly, appropriate apparatus may be applied to determine the fastness of a joint in woodwork, the tensile strength of a metal product, and the like.

Toops (*28*) has proposed a method which employs code numbers to objectify and facilitate the scoring of the product of performance tests. Through the use of this rather simple but ingenious procedure, parts of the product are identified by numbers and letters, and the scoring becomes a simple clerical operation requiring little or no mechanical skill or understanding on the part of the proctor or recorder. Figure 54 illustrates

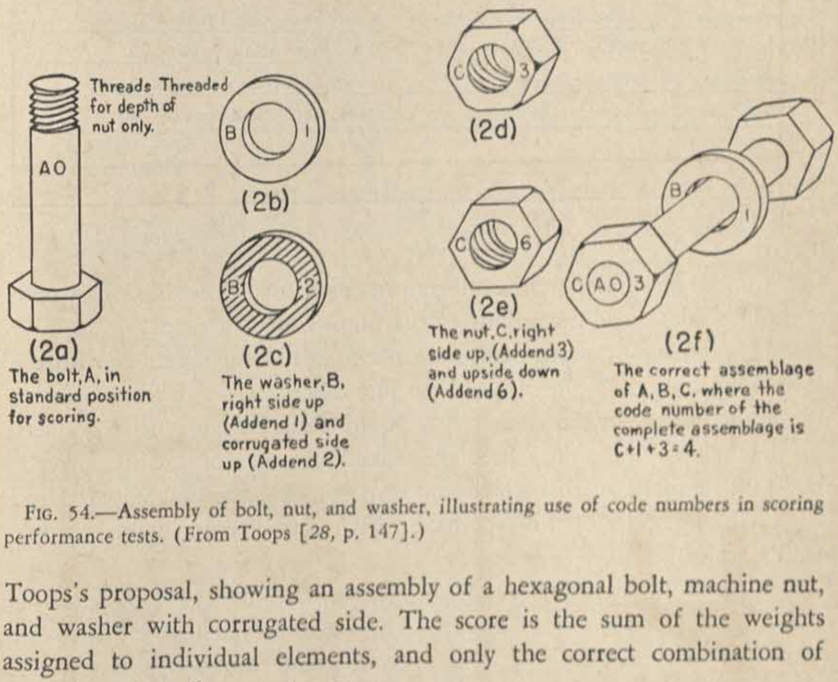


FIG. 54.—Assembly of bolt, nut, and washer, illustrating use of code numbers in scoring performance tests. (From Toops [*28*, p. 147].)

Toops's proposal, showing an assembly of a hexagonal bolt, machine nut, and washer with corrugated side. The score is the sum of the weights assigned to individual elements, and only the correct combination of operations gives the correct sum.

Although it is possible to measure many products of performance objectively, more subjective methods must be resorted to in some cases. In the judgment of handwriting or freehand drawing, for example, there is no pattern or gauge or scale that can be used to yield a direct measurement of goodness or satisfactoriness of the product. In such situations, rating devices of various sorts may be employed. While the estimates derived from such methods are likely to be less reliable than those of

more objective techniques, they frequently can be refined to a point where their use can be defended.

The Minnesota Food Score Cards (*5*) were devised to measure the quality of foods that had been prepared in home economics classes. The procedure is not highly refined from a psychophysical standpoint. In use, a score or rating is simply assigned to each of various characteristics or qualities of the food prepared. Figure 55 shows the score card used in rating the cooking of bacon.

BACON **2**

	1	2	3	Score
Appearance.....	1. Curled or has decided humps		Fairly straight and flat	1.
Color....................	2. Too light or dark brown		Even, light golden brown	2.
Tenderness.......	3. Brittle or tough		Crisp, easily cut with fork	3.
Flavor...................	4. Burned, acrid, raw		Mild, meaty, well done	4.

SCORE ______

FIG. 55.—Score card used in rating the cooking of bacon. (From Brown [*5*].)

A somewhat similar point-scale rating form for "fastening" in woodworking, as described by Adkins (*1*), is shown in Figure 56.

Some products of performance are not as readily analyzed and broken down into units or segments as those just described. In order to achieve maximum objectivity and reliability in the judgment of such products, a series of samples of the product, ranked in order of quality, may be used as a standard against which any given product can be compared and scored. Rating scales of this type have been referred to as "graded sample quality scales."

Such psychophysical methods as the *method of equal appearing intervals* and the *paired comparisons method* may be employed to establish scale values for samples of products of varying grades of quality. The judgment of the quality of a particular product may then be made by comparing the product in question with the graded samples and assigning the value, or score, of the sample that the product most closely resembles.

One of the most familiar examples of such a scale of samples of graded quality is that which has been used for assessing the quality of handwriting. An example of the Ayres Handwriting Scale is shown in Figure 57 (*2*).

(*a*) Nails:

(1) Straightness	1	2	3	4	5	6	7	8	9	10
	Are nails driven straight, heads square with wood, no evidence of bending?									
(2) Hammer marks	1	2	3	4	5	6	7	8	9	10
	Is wood free of hammer marks around nails?									
(3) Splitting	1	2	3	4	5	6	7	8	9	10
	Is wood free of splits radiating from nail holes?									
(4) Depth	1	2	3	4	5	6	7	8	9	10
	Are depths of nails uniform and of pleasing appearance?									
(5) Spacing	1	2	3	4	5	6	7	8	9	10
	Are nails spaced too close or too far apart?									
(6) Utility	1	2	3	4	5	6	7	8	9	10
	Will the nails hold?									

(*b*) Screws:

(1) Slots	1	2	3	4	5	6	7	8	9	10
	Are slots free of splitting and other evidence of driving strains?									
(2) Straightness	1	2	3	4	5	6	7	8	9	10
	Are screws straight, heads parallel with surface?									
(3) Splitting	1	2	3	4	5	6	7	8	9	10
	Is wood free of splits in the area of screws?									
(4) Screw driver marks	1	2	3	4	5	6	7	8	9	10
	Is wood free of screw driver marks near screws?									
(5) Countersinking	1	2	3	4	5	6	7	8	9	10
	Is countersinking neat and of satisfactory depth?									
(6) Spacing	1	2	3	4	5	6	7	8	9	10
	Are screws spaced too close or too far apart?									
(7) Utility	1	2	3	4	5	6	7	8	9	10
	Will the screws hold?									

FIG. 56.—Point-scale rating form for "fastening" in woodworking. (From Adkins [*1*, p. 231].)

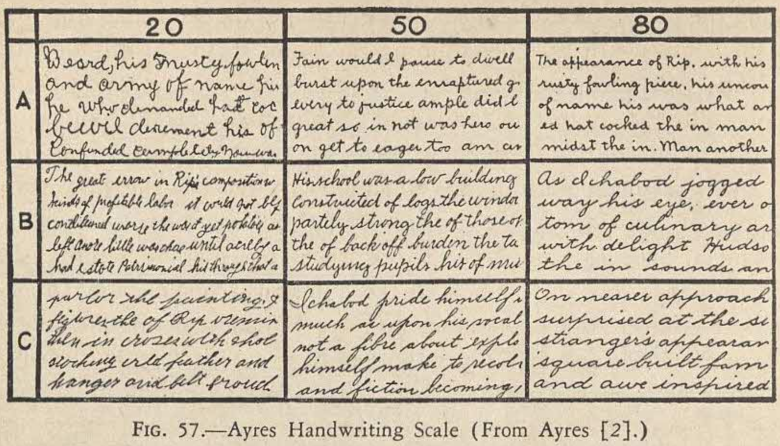


FIG. 57.—Ayres Handwriting Scale (From Ayres [2].)

Figure 58 shows a graded sample quality scale used in determining the excellence of Western Union splices made by electricians. Scale values have been assigned to each of the samples ranging from excellent products to those that are very poor (*19*).

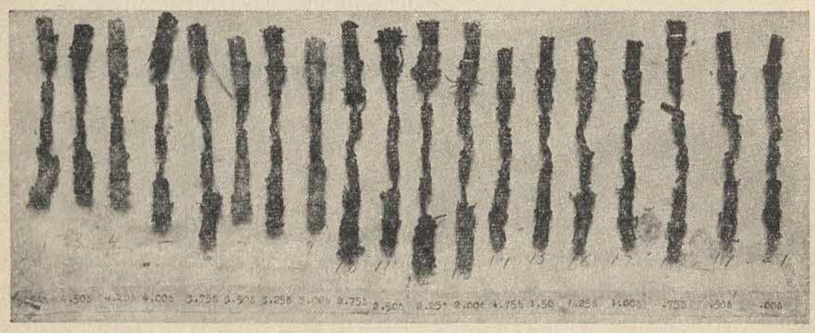


FIG. 58.—Graded sample quality scale for judging the excellence of Western Union splices made by electricians. (From Paterson [*19*, p. 190].)

The graded sample quality scale is relatively easy to devise, although some acquaintance with psychophysical methods is necessary. A large number of samples of the product are assembled, and these are ranked in order of quality by experts. Samples may then be selected in accordance with the judgments of the experts and with regard to clear-cut and consistent differences in judged quality.

The problem of scoring standards.—Ordinarily, it is desirable to employ a scoring procedure which will satisfactorily distribute (or spread out) the scores of examinees with respect to the performance being measured.

In using the results of a test which thus distributes the scores, an important problem is always that of determining the point in the distribution that will define satisfactory performance—in other words, the "passing point." It is one thing to say that an individual's performance netted him a given number of points or that these points were equivalent to a certain percentile rank, and sometimes quite another to attempt to translate the score points into terms of "satisfactory" or "unsatisfactory" performance.

Setting a "passing point" for performance tests may be accomplished by relating the score distribution to the criterion in some cases, but this often is an uncertain undertaking. Just what total score on the Rating Form for Fastening shown in Figure 56 represents satisfactory performance? Would the performance an individual who received a rating of 10 on all of the "screws" items except items 3 and 7, and who hopelessly split the wood into which the screw was being driven, be considered satisfactory? Would his raw score of 50 represent approximately the same level of attainment as the score of another individual who received ratings of 7 on each of the seven items? Usually one must resort to the judgment of competent persons and arbitrarily decide upon the pass-fail point or other defined points of a score distribution.

In many situations in which performance tests are used, the "minimum essentials" of the performance may be considered in establishing the satisfactory-unsatisfactory point. Such a test might consist of selected essential operations, or qualities, all of which must be properly demonstrated in order for the performance to be satisfactory. During the war, for example, tests of this sort were developed for use at one of the amphibious training bases, where it was necessary that every trainee be able to recognize beachmarker signals indicating where certain types of cargo were to be unloaded, signals identifying various types of landing craft, and the like. A minimum essentials test would be appropriately used in situations involving relatively simple operations without mastery of all of which it would be dangerous or impossible for a person to carry out the activity in question.

Reliability

In scoring any test, the judgment of an observer is to some degree involved, even though the scoring may consist only of counting the dots that appear through a scoring stencil or reading the dial of a test-scoring machine. In performance testing, the role of the observer may become considerably enhanced in importance. Not everyone is qualified, for example, to grade the performance of a piano concerto or to judge the

skill of a trainee in using a lathe. If the over-all reliability of a performance test is found to be high, one need not be concerned about the reliability of observer judgment; but if the reliability is low, it is necessary to study the relative contributions to a reliability coefficient of (1) consistency of performance of the testee and (2) consistency of judgments of the observer who is making the evaluations.

Data for studying observer reliability may easily be obtained, if some *product* of the performance such as a theme or a hand-tool project is to be rated, merely by having the same set of products independently judged by several observers. If the performance itself is to be judged, it is ordinarily possible to arrange for several observers to view the performance simultaneously and make independent judgments. If the intercorrelations of the several sets of ratings are on the average high (say, .90), reliability of the observers may be said to be satisfactory. This method also permits the identification of the poorest judge on the basis of lower correlations of his ratings with those of other judges. In the event that the intercorrelations are on the average low, steps must be taken to improve the reliability of judging the performance.

The nature of the steps to be taken will, of course, depend on the characteristics of the testing situation. (1) In some instances, selection of better-trained observers might be sufficient. (2) Even with well-qualified observers, it might also be necessary to give special training with respect to definitions of the characteristics to be rated and of the items on the scale used in rating in order to insure agreement among judges on these points. (3) Designing a rating sheet such that various aspects of the performance are objectively defined and independently judged and on which instructions for use are clearly stated may improve consistency of rating. (4) Development of instruments for objectively measuring the products may bring about the desired improvement in reliability.

For example, in one shop training situation it was found that the correlations among raters in judging metal objects constructed to certain specifications by trainees varied from .11 to .55. When a set of simple taper gauges was constructed for use in measuring the products and when the raters were instructed in their use, the correlations for the same judges increased to .93 and .94. When the same raters repeated the measurements after an interval of ten days, the correlation of first ratings with second ratings was found to be .97.

It is possible to study the *reliability of performance* (as distinguished from judging performance) only when the reliability of judging performance has been shown to be adequate. When this condition has been

satisfactorily met, the special problems having to do with measuring the consistency of performance of the examinee should be considered.

When a performance test is made up of a series of tasks involving, let us say, the disassembly and assembly of a machine, its adjustment and operation, and the analysis of defects in the machine, a high degree of specificity is often found in ability to perform the various tasks, particularly in a training situation. An individual who can do one task well may do poorly on other tasks. Such specificity may indicate a real lack of generalized ability in the area of mechanical manipulation, or it may merely reflect irregularities in the training program. Whatever the cause, such a situation would appear to show low test reliability when reliability is measured by correlating scores on one set of tasks with scores on a second set of tasks (split-half or alternate-form reliability). It may sometimes be necessary to make a special investigation to determine whether the low reliability is the fault of the test or the fault of a training situation which does not consistently give the same amount and quality of instruction to each student on each part of the curriculum. In such cases the use of test-retest reliability may be appropriate.

However, test-retest reliability also has distinct limitations in certain situations. In a training situation it may be found that a test-retest type of reliability may not be feasible because participation in the first test situation may constitute a significant amount of additional practice for some of the trainees, but not for others. The test-retest correlation may be seriously affected by such circumstances.

Variation in condition of the equipment used in a performance test is still another source of unreliability which must be taken into account and guarded against, particularly when complex operations, such as difficult assembly jobs or precise adjustments, are involved.

Perhaps the essential difficulty in the development of performance tests of high reliability is that relatively few tasks or items are likely to be involved in the testing. In cases where many simple short operations are required, as in operating a typewriter, it is not difficult to build tests of satisfactory reliability; but in any operation involving relatively long, complex tasks, such as in the assembly of an engine or the construction of a woodwork sample, the performance test is likely to be of a lower reliability unless it is an exceptionally lengthy one.

Steps in Developing a Performance Test

1. Making a job analysis

The first step in the development of a performance test is to make

a very careful study of the specific skills and abilities involved in activities the test is intended to measure. Such a study might best culminate in a formal job analysis report. Methods of job analysis have been adequately described elsewhere, and it will not be necessary to discuss the procedures here. If at all feasible, the test constructor might well be advised to learn the job himself. He may thus come to understand better some of the more subtle aspects of the performance which are difficult to convey in words. The usual observational methods involved in making the job analysis might also be supplemented by such devices as time-and-motion studies and analyses of causes of failure on the job.

If the performance test is to be employed in a training situation, it usually is desirable to go beyond the stated objectives of the curriculum, and study the job for which training is being given. It is quite possible that the curriculum is in some respects unrealistic, and that tests which would adequately assess school achievement might not be as closely related to job success as would be possible or desirable. In developing tests of school achievement, it is entirely possible that contributions can thus be made which will lead to the improvement of the training situation.

2. Selecting tasks to represent the job

The next step in performance test construction is to determine which of the operations or skills are to be tested among those described in the job analysis, and what specific tasks are to be chosen to represent those skills. In general, as wide a variety of specific tasks should be included in the test as is feasible from the standpoint of such practical considerations as amount of time and equipment available. Special care should probably be taken to insure that any skills which are commonly responsible for failures are included. Short tasks usually are preferable to long ones, since a larger number of items can be included in the test and higher reliability probably secured.

After it is decided what abilities are to be tested, it is necessary to determine whether the performance of the task itself (performance in process) or some product of that performance should be evaluated. Sometimes the solution is obvious; if a test for truck drivers is to be developed, there is no product to be judged and the only possibility is to observe the performance itself. On the other hand, the product may be of primary importance and the process by which that object is produced of no particular consequence. In a test of English composition where the product is an essay, one would not ordinarily be concerned with the time required, the revising and rewriting indulged in, or the frustrations and blocks

experienced during the preparation of the essay. In still other situations both the performance in process and the characteristics of the product may be of importance; for a given type of performance it might be desirable to note whether or not safety precautions were observed, if procedures likely to damage equipment were employed, or if wasteful errors had been made (even though later corrected), as well as to rate the product on its important characteristics.

In the event that a product is to be evaluated, it is necessary to decide what workpieces or job samples will be produced. The specifications of the product must be described in sufficient detail so that variability among products of different examinees is a function of differences in ability of testees rather than differences in interpretations of the job specifications. In the measurement of proficiency in use of hand tools, a test situation might be developed which would require the use of layout tools, hack saw, and files. On completion of the workpiece, the product could be measured and inspected and ratings made to describe success in using these tools. A still better method might be to rate the product at the end of each of various stages in its production—after layout, sawing, rough filing, etc. Such a procedure probably would yield more reliable scores and would furnish information which might have greater diagnostic value. Other test situations might similarly be arranged to yield measures of proficiency in the use of other hand tools.

In setting up the job specifications for a test, it should be kept in mind that *time* is an important consideration in performance testing. If a testee is required to spend a large proportion of his testing time in routine or repetitive operations, fewer items can be included in the test and its reliability will necessarily suffer. Ordinarily, for example, a half-inch hack saw cut would be as good as one inch for testing purposes and would take roughly half the time, perhaps allowing an opportunity for adding another test item.

Similar considerations are involved in selecting tasks where the performance in process is to be observed. As wide as possible a coverage of the basic skills revealed by the job analysis should be sought, including particularly any tasks which give opportunity for errors which have been found to be common causes of failure. Testing time should be used to the fullest possible extent for performing the crucial or difficult aspects of the job, rather than routine operations.

3. Developing the rating form

Having selected a series of tasks or qualities to be included in the

performance test, the next procedure is to determine the features of the performance, or of the product, which are to be rated, and to devise suitable rating forms.

The rating form or record sheet may consist merely of a listing in correct order of the operations which must be carried out to perform the job assigned, with space to check whether or not each operation was correctly performed. It is possible also to permit the judge to rate the quality of the performance of each operation, but in many situations it is preferable to reduce the scoring to a yes-no check on each characteristic of the performance considered important, making sure that the performance description is sufficiently unambiguous as to permit high observer agreement.

The performance descriptions employed may be given differential weighting if it is desired to increase the variance in total score contributed by certain operations or qualities.

If time required to perform the task is important, this may also be recorded and used in the scoring of the test.

When the product of the performance is to be rated, a list of the characteristics which serve to differentiate a good from a poor product should be compiled. Some of these characteristics may be amenable to direct measurement with a gauge or other measuring device; others may require a more subjective judgment. The accuracy of such judgments may be increased by the use of graphic or descriptive rating scales, or by comparison with a series of products which have been scaled with respect to the attribute under consideration. These topics have been discussed in some detail in an earlier section of this chapter.

4. *Surveying the practical limitations*

Before developing an "operating plan" for administering a performance test, due account should be taken of various factors in the situation which impose limitations upon the "ideal" procedure which might otherwise be outlined. These restrictions are likely to be related to such factors as amount of time, equipment, and personnel available.

One of the first considerations at this point is the amount of time which can reasonably be devoted to performance testing. This decision will hinge on such questions as the importance of the task, the probable validity of the test, the number of applicants, and the cost of equipment and personnel needed for the test administration. In a school situation it would perhaps be necessary also to estimate what proportion of the total time scheduled for instruction may be devoted to performance testing.

Again, the conclusion would depend upon the particular situation. In certain kinds of instruction such as typewriting, it might be advisable to spend as much as three-fourths of the time in performance testing, that is, in timed trials. In other areas of study the proportion obviously would be much smaller.

Some investigation also will be necessary to determine how much equipment may be available in relation to the number of candidates to be tested; and how many persons may be available who will be qualified to administer the test or to make the required evaluations of performance. The operating plan will be prepared in the light of such considerations as these.

5. Developing the tentative "operating plan"

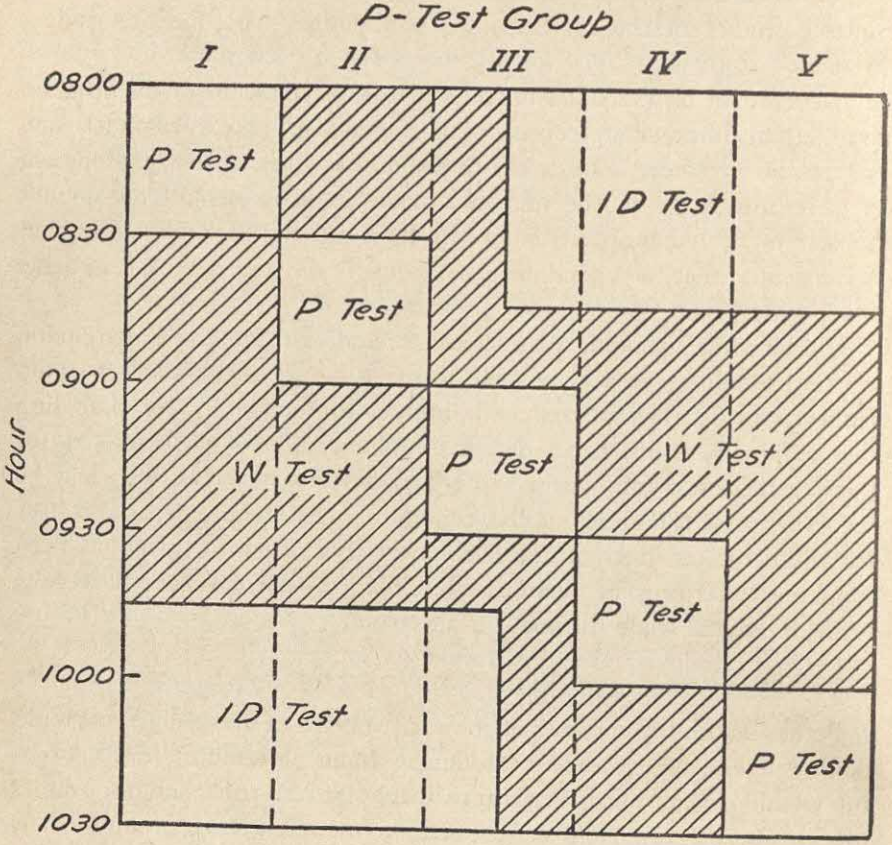
After having analyzed the job, selected tasks to represent the job in the test situation, developed trial forms for rating or recording performance, and surveyed the practical limitations imposed by considerations such as time, personnel, and equipment, the next procedure in developing a performance test is to organize the data and materials—to formulate an operating plan. This plan will be tentative in nature at the beginning and will become more crystallized as the construction of the test progresses. Numerous trial runs of portions or all of the test may be required to establish suitable time allowances, adequate instructions, feasible means of judging certain aspects of performance, and the like.

In developing the operating plan, the exercise of ingenuity may sometimes enable one to overcome obstacles related to limitations in time, personnel, and equipment, which might otherwise reduce the proposed testing program to dimensions which would be fatal to its success. Where the amount of equipment is too small to permit testing of more than a few examinees at one time, it may be possible to break down each piece of equipment into subassemblies, permitting one man to work at each subassembly. Since it is not always essential that the jobs be done in any particular order, examinees can rotate from one subassembly to another, and thus a larger number may be tested simultaneously than would otherwise be possible.

The administration of several types of tests also can sometimes be carried out simultaneously in order to make maximum use of the testing time and equipment available. For example, a performance test might be given to successive small groups (of a size that can be accommodated by the amount of available equipment), drawing them out of a larger group which is occupied with, say, a paper-and-pencil test. In some school

situations as many as three tests have been simultaneously administered, following a scheme which is diagrammatically shown in Figure 59 (*24*).

If it is necessary to make on-the-spot evaluations of the performance



This procedure permits all the men to take all three tests within a period of three hours without conflict. Maximum use is made of the available testing time. It is necessary that all groups except the first and last group taking the performance test be interrupted once in their written test; this is not a serious objection if the written examination is of the short-answer or objective type.

The diagram assumes that 10 men can be tested at once on the performance test. The system can be used just as easily with P-test groups of 8 or 9 men. In these cases, the last P-test group would contain fewer men than the other groups, if there are 50 men in the class.

Some care must be taken to prevent students from talking with each other about the written test items while going to and from the performance test.

FIG. 59.—Diagram illustrating the method of coordinating the administration of identification, performance, and written tests. (From Stuit [*24*, p. 475].)

exhibited during the test (performance in process), the lack of well-trained judges may restrict the number of examinees who can be tested at one time. This difficulty can sometimes be overcome by preparing rating forms that are adaptable for personnel who lack the full amount of

training. Advanced students, for example, can sometimes be used to judge the performance of elementary students, provided they are adequately supervised and given a sufficient amount of special instructions for the task. Such a procedure should, of course, be justified by making suitable studies of the agreement of such judges with the best judges available.

There are, of course, likely to be limitations in the scope of the program even after all possible economies in the use of time, personnel, and equipment have been considered. In the initial stages of the development of a testing program, the test constructor probably should not permit himself to be handicapped unduly by these practical considerations; but it is essential that the operating plan as finally developed be a reasonable and practical one.

The tentative operating plan should provide as complete a description of the procedures as it is possible to make before a full-scale trial run has been made. The plan should include directions (1) for preparing the equipment or other materials for administration of the test, (2) for training the test administrator and other assistants or judges who may be involved, (3) for checking the condition of the equipment, and (4) for the actual conduct of the test. After this operating plan has been subjected to scrutiny by qualified critics, and suitable revisions have been made, the test is ready for its full-scale tryout.

6. Tryout and revision of the test

Before its use for purposes of measurement the performance test should be thoroughly studied under realistic conditions of administration. A trial run should be arranged, permitting the test to be tried out on samples of examinees drawn from the population for which the test is ultimately intended. The samples used in the tryout should be as representative as possible, and large enough to yield stable statistical results.

During the tryout, the tentative operating plan should be carefully followed in all details. Arrangements should be made to have the examinees' behavior and all procedures observed in detail to detect possible shortcomings and to suggest improvements. Special precautions should be taken to see that the equipment is adequately prepared and the test administrators thoroughly trained; otherwise the usefulness of the test cannot be properly determined.

The value of the trial administration for confirming the usefulness of the test, or for suggesting improvements, depends upon the careful observation of the procedures and the keeping of complete records. The

test scores themselves constitute an important part of these records, but they should be supplemented by notes of undesirable aspects of the testing and unexpected behavior of the examinees as they are observed.

If judgments of the performance in process are involved in the test, provision should be made for two or more judges to make independent ratings of each candidate's performance, so that the reliability of the test may be studied.

The difficulty of the test situations or items, as revealed by the tryout, will be of particular interest. If the test is of the minimum essentials type and the group tested has been well trained, most of the items or tasks probably will be successfully passed by a large proportion of the candidates. If it is not a minimum essentials test, the proportion passing each item ordinarily will be smaller. Certain items may be found which are passed by very few of the candidates. Before discarding these items on the grounds that they are "too hard," the test constructor should be sure that their elimination will not harm the test from the standpoint of validity. In the performance testing in Navy service schools, it occasionally was found that instructors had failed to teach certain skills, although that aspect of the training had been assumed to be satisfactory. Items which are failed by a large proportion of the candidates may reveal deficiencies in training rather than in the test itself. On the other hand, it may be that the test item requires an unreasonably high level of proficiency for the candidates being tested, and the item then should be revised or eliminated. Still another possibility, if the item involves evaluation of the quality of performance by a judge, is that the standards of judgment involved in the rating are unrealistic, and that the descriptive rating scale needs revision or the judges should be given more adequate instructions. In general, the principles of item selection presented in chapter 9 apply equally to performance tests.

In situations where subjective ratings of performance are involved, the agreement of the various judges who have independently made observations should be studied. This probably can best be accomplished by computing correlations for each pair of judges for the ratings of performance on each item or task. Ideally more than two judges should be used, making it possible to compare the judges' estimates with respect to their agreement with each other. If the correlation is high for certain judges, it would seem to establish the possibility of achieving reliable evaluation of performance, even though correlations were low for other judges. If the reliability of judgment as measured by correlations between judges is consistently low, then steps must be taken to improve reliability.

Additional experimentation may be necessary to discover what steps should be taken—improve the training of the judges, revise the rating form, or change the nature of the task performed by the examinees.

Study of the internal consistency of a performance test may or may not be relevant to the problem of the test's reliability. It would probably be useful in any case to study the tryout data to determine the extent to which success on one task is associated with success on other tasks. But high internal consistency is not to be expected for certain types of tests, particularly if they are of a composite nature or if they are used in a training situation. If a split-half correlation is high, it does not necessarily mean that the test is highly reliable; it may be desirable to attempt to obtain other evidence of reliability. A test-retest correlation, for example, might furnish a better estimate of the reliability of the tests, although if the test is used in a training program, the test itself might furnish a significant amount of additional practice in the case of some trainees and thus spuriously affect the obtained reliability coefficient. In general, the principles and procedures described in chapter 15, "Reliability," apply equally to performance and other types of tests.

A study of the evidence bearing on the question of reliability may suggest improvements in the test or in a training program in which it is used. Uneven learning, as shown by trainees who do well in some tasks but poorly in others, may direct attention to the training program and reveal that some students are given better training in certain areas than others. If the unreliability is clearly a function of the test itself, and not due to uneven preparation of trainees or to unreliability of ratings made by the judges, then some modification of the test is necessary. The most obvious solution of the problem would be to lengthen the test. If this is impossible, then an alternative might be to increase the number of scorable units of performance in the time available.

Evidence as to the validity of the performance may be particularly difficult to obtain. If the test is used in a training program, it may be possible to correlate test scores with other measures of achievement such as course grades. But such correlations must be interpreted with caution. For course grades often reflect verbal ability or mathematical ability more than skill in performance; so low correlations will not necessarily indicate low validity. Another possibility would be to select the highest and lowest examinees in the group on the basis of performance test scores, and ascertain from instructors or supervisors whether or not in their opinion these individuals were clearly in different categories of achievement. Such a rough type of validation should reveal gross failure of

the test to measure what it is supposed to measure. (For a more adequate discussion of the problem of validity, see chapter 16.)

7. *Preparation of directions for administration and use of the test*

After the test has gone through at least one trial administration and revision, and when the test is judged to be ready for use, a complete manual of directions should be prepared. The importance of such a detailed manual is often underestimated; it is easy to believe that instructions will be remembered by proctors and that the details regarding preparation of equipment and the like will be carried out uniformly from time to time. People do forget, however, and personnel may change. A complete, detailed description of the procedures in manual form will help to insure uniformity of conditions and procedures from one test administration to the next.

The preparation of a manual of directions is considered in general in chapter 10. The following points should be especially considered in the performance testing situation.

1. *Preparations for administering the test.*—Under this heading should be included specific directions for: (1) training proctors or other assistants who may assist in giving the test or in judging the performance of the candidates; (2) the arrangement of equipment and the materials which are to be used in testing; (3) preparing the equipment for testing, by making whatever adjustments or maladjustments that may be necessary, introducing defects or misassembling parts, if required, and checking on deterioration of equipment; (4) providing for a thoroughgoing practice session with proctors or other examining assistants; and (5) arranging for the coordination of performance testing with other testing sessions.

2. *Conduct of the test.*—Under this heading should be discussed the detailed directions for administering the test, including instructions for giving directions to the examinees, for maintaining standard conditions, for timing the test and recording time and other records of performance, and for resetting the apparatus for the next group of examinees.

3. *Scoring the test.*—Here should be outlined the directions for grading performance or a product, for recording the scores on appropriate forms, for checking the scoring, and for converting the raw scores to standard scores, percentiles, or other types of measures.

Many paper-and-pencil tests are essentially self-administering. Performance tests very seldom can be so considered. The control of testing conditions is basic to the success of a procedure, and care must be exercised

constantly to assure uniform administration. In some situations, such as radio code receiving and stenography (*23*), the possibility of uniform testing conditions may be increased through the use of recordings which permit a high degree of control over the presentation of test materials.

Performance tests in many respects may be considered laboratory experiments. As such, they require the same insight in designing and the same vigilance in administering.

Selected References

1. ADKINS, D. C., *et al. Construction and Analysis of Achievement Tests.* Washington: Government Printing Office, 1948. 292 pp.
2. AYRES, L. P. *A Scale for Measuring the Quality of Handwriting of School Children.* New York: Russell Sage Foundation, 1915. 16 pp.
3. BARNES, R. M., and AMRINE, H. T. "The Effect of Practice on Various Elements Used in Screw-Driver Work," *Journal of Applied Psychology,* **26**: 197–209, 1942.
4. BLUM, M. L. "Selection of Sewing Machine Operators," *Journal of Applied Psychology,* **27**: 35–40, 1943.
5. BROWN, C. M., *et al. Minnesota Food Score Cards.* Minneapolis: University of Minnesota Press, 1946.
6. CHAPMAN, J. C. *Trade Tests.* New York: Henry Holt, 1921. 431 pp.
7. CURETON, T. K.; BOOKWALTER, K. W.; GLASSOW, R.; and MCCORMICK, H. G. "The Measurement of Understanding in Physical Education," *The Measurement of Understanding: Forty-fifth Yearbook of the National Society for the Study of Education,* Part I. Chicago: National Society for the Study of Education, 1946. Chap. 12, pp. 232–52.
8. DESILVA, H. R., and CHANNELL, R. "Driver Clinics in the Field," *Journal of Applied Psychology,* **22**: 59–69, 1938.
9. FREEMAN, F. N. *Freeman Chart for Diagnosing Faults in Handwriting.* Boston: Houghton Mifflin, 1914.
10. FRUTCHEY, F. P.; DEYOE, G. P.; and LATHROP, F. W. "The Measurement of Understanding in Agriculture," *The Measurement of Understanding, Forty-fifth Yearbook of the National Society for the Study of Education,* Part I. Chicago: National Society for the Study of Education, 1946. Chap. 14, pp. 270–80.
11. HAY, E. N. "Predicting Success in Machine Bookkeeping," *Journal of Applied Psychology,* **27**: 483–93, 1943.
12. HULL, C. L. *Aptitude Testing.* Yonkers, N.Y.: World Book Co., 1928. 535 pp.
13. LAWSHE, C. H. *Principles of Personnel Testing.* New York: McGraw-Hill, 1948. 227 pp.
14. LAWSHE, C. H., and TIFFIN, J. "The Accuracy of Precision Instrument Measurement in Industrial Inspection," *Journal of Applied Psychology,* **29**: 413–19, 1945.
15. LINDAHL, L. G. "Movement Analysis as an Industrial Training Method," *Journal of Applied Psychology,* **29**: 420–36, 1945.
16. MCCLOY, C. H. *Tests and Measurements in Health and Physical Education.* New York: Crofts & Co., 1942. 412 pp.
17. MCPHERSON, M. W. "A Method of Objectively Measuring Shop Performance," *Journal of Applied Psychology,* **29**: 22–26, 1945.
18. NEWKIRK, L. V., and GREENE, H. A. *Tests and Measurements in Industrial Education.* New York: John Wiley & Sons, 1935. 253 pp.
19. PATERSON, D. G.; ELLIOTT, R. M.; *et al. Minnesota Mechanical Ability Tests.* Minneapolis: University of Minnesota Press, 1930. 560 pp.
20. PATTEN, E. F. "An Experiment in Testing Engine Lathe Aptitude," *Journal of Applied Psychology,* 1923, **7**: 16–29.
21. PROFFITT, M. M.; ERICSON, E. E.; and NEWKIRK, L. V. "The Measurement of Understanding in Industrial Arts," *The Measurement of Understanding, Forty-fifth Yearbook of the National Society for the Study of Education,* Part I. Chicago: National Society for the Study of Education, 1946. Chap. 16, pp. 302–20.

22. SCOTT, M. G., and FRENCH, E. *Better Teaching Through Testing.* New York: A. S. Barnes Co., 1945. 247 pp.
23. SEASHORE, H., and BENNETT, G. K. *Stenographic Proficiency Tests.* New York: Psychological Corporation, 1946.
24. STUIT, D. B. (ed.). *Personnel Research and Test Development in the Bureau of Naval Personnel.* Princeton, N.J.: Princeton University Press, 1947.
25. THORNDIKE, R. L. *Personnel Selection.* New York: John Wiley & Sons, 1949. 358 pp.
26. TIFFIN, J. *Industrial Psychology.* New York: Prentice-Hall, 1947. 554 pp.
27. TIFFIN, J., and GREENLY, R. J. "Experiments in the Operation of a Punch Press," *Journal of Applied Psychology,* **23**: 450–60, 1939.
28. TOOPS, H. A. "Code Numbers as a Means of Scoring Group-Administered Performance Test Products," *Journal of Applied Psychology,* **26**: 136–50, 1942.
29. ———. *Trade Tests in Education.* New York: Teachers College, Columbia University, 1921. 118 pp.
30. VITELES, M. S. *Industrial Psychology.* New York: W. W. Norton & Co., 1932. 652 pp.

13. The Essay Type of Examination

By John M. Stalnaker
Association of American Medical Colleges

Collaborators: E. S. Noyes, *Yale University;* Verner Sims, *University of Alabama*

Under the heading of essay examination questions, many different question types are included, ranging from a single three-hour question such as "Discuss Milton's poetry," to a short-answer question which may be answered completely and correctly in a sentence or two, such as, "Give one example of the democratic thesis in the play *Ruy Blas.*" For the purposes of this chapter—concerned with tests used for determining relative achievement in a course or field of subject matter—the essay question is defined as a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject. The most significant features of the essay question are the freedom of response allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerks to check, but even an expert cannot usually classify a response as categorically right or wrong. Rather, there are different degrees of quality or merit which can be recognized.

A definition of a functional type has been offered recently by Sims (*4*, page 17): "The essay examination is a relatively free and extended written response to a problematic situation or situations (question or questions), which intentionally or unintentionally reveals information regarding the structure, dynamics and functioning of the student's mental life as it has been modified by a particular set of learning experiences." This definition can in general be applied to the essay question as discussed in this chapter, except for certain differences in emphasis. The essay question, as the term is used here, need not require an extended response. Indeed, the use of questions calling for brief responses will be urged for measuring certain types of outcomes. An essay question, it will be shown,

can be framed which requires only a brief response and yet which allows the examinee considerable freedom and elicits a response which reveals pertinent information about the candidate's learning.

In his article, Sims considers the essay question as a projective technique, a technique in which a problematic situation is presented to an examinee in an effort to force him to "project" his personality, as it has been modified by some particular learning, into the answer, and to make choices in terms of his own experiences and sense of values. In responding to any test question, regardless of form, the examinee brings his personality into play and uses his own experience and sense of values to some extent. But the extent to which this is true varies greatly from question to question. Some questions allow great freedom; for example, "What has this course in philosophy meant to you?" The problem of the sincerity of the response is not easily solved. Other questions place more emphasis on limited facts and reasoning learned in a course; for example, "Differentiate between the term dialectic as used by Plato and as used by Hegel." In this question, the examinee may not have great freedom to express his personality, but the answers to a series of questions like it may do more to reveal reliably a knowledge of the subject matter which is felt by the instructor to be important. The projective technique is not a suitable, dependable method, in the present stage of its development, to determine the attainment of many of the usual objectives in the school situation, although consideration of the essay question as a projective technique may bring into focus certain important characteristics of the essay question and perhaps warns against certain dangers.

The essay question as discussed here includes "problematic situations," but in the broadest sense of that term. Essay questions may include such humdrum questions as "Why is the bulk of concentrated sulfuric acid made by the 'contact process'?" Such questions may be, and usually are, largely questions of memory. Yet, unless the correct answer to the specific question has been taught, even these questions involve understanding which goes beyond rote memory.

The latter half of Sims's definition, as its author has observed, can be applied to any examination question—essay, objective, or performance. All educational measurements are generally intended to elicit information regarding the structure, dynamics, and functioning of the student's mental life as it has been modified by a particular set of learning experiences. The special problem in the case of the achievement test is to obtain information which is reliable and pertinent, and to do so efficiently.

The Essay Question in Current Achievement Testing

The essay question continues to be the most widely used test form in achievement testing today. It has remained almost untouched by the widespread and continued criticism of the experts in educational measurement. The reasons for this situation are not hard to find. The most widely used test is the one prepared, administered, graded, and interpreted by the classroom teacher, and intended to be used only with pupils in a single course taught by the teacher preparing the test. It is used for a variety of purposes: to motivate the pupils, to help in determining their achievement in the course, and to encourage proper study habits. All over the world, thousands of teachers, most of whom are unaware of the modern techniques in testing described in this volume, are preparing essay questions to which pupils from elementary school to the graduate level write out answers rather than make marks on an answer sheet indicating the choice of one of several fixed responses. The teacher, who, without experience or technical training, has endeavored to use objective tests for his own class has often been dissatisfied with the results, and rightly so. The preparation of an objective test almost always requires more time as well as special skill, and the interpretation of the results of such a test when restricted to a single class is difficult at best. Nearly every teacher, however, considers himself fully capable of setting a satisfactory essay test in his own course or subject, and of reading the answers at least to his own satisfaction. The grades assigned to the papers are usually based in part on what the student has written and in part on what the teacher, from other knowledge of the candidate, believes he meant by what he wrote. Norms, correlations, reader or test reliability, and validity do not worry the teacher. If he knows what they are, he ignores them in dealing with his own classroom situation. The popularity of the essay question should not, therefore, be misinterpreted to indicate that it is the most suitable test form for many purposes, that it is in a "healthy" condition, or that improvements are not needed.

Conversely, the diminishing use of essay questions by the testing agencies concerned with educational measurement does not establish that the essay question has no unique or important measurement values. The essay question is being used less and less in achievement tests developed for wider use than by one teacher in a single classroom. Fixed-response tests, frequently standardized so that scores of pupils in one class may be compared with those of a defined larger population, are enjoying a growing popularity in the elementary and secondary school, and even

to some extent now in the colleges. The College Entrance Examination Board, for example, has given up the essay question completely except in the field of English composition, and even here the use of the essay question is shrinking; yet the College Board tests are being used more extensively than ever before. The Regents' Tests of New York State are becoming increasingly of the short-answer type. The Educational Testing Service uses the recognition type of objective test almost exclusively. State-wide testing agencies generally restrict their tests to objective items. The Army, the Navy, and Civil Service are all turning to objective tests for the purpose of classification and measurement of achievement. Even the Department of State in its selection of personnel for the foreign service—a selection from among college graduates—is using an increasingly large proportion of fixed-response items. The tests of these groups, and even many of the tests used in college courses with large student enrollment, are prepared by test technicians in collaboration with teachers of the subject matter concerned.

The fact that these specialists in measurement have devised the objective item to overcome some of the weaknesses which they believe to be inherent in the essay question cannot be interpreted as meaning that the essay examination has no useful place in educational measurement. Their criticisms, such as the high cost of reading essay papers, the difficulty of obtaining reliable readings, the problem of securing an adequate and dependable sample of student behavior, the difficulty of obtaining a wide range of scores, and the task of framing questions so that all candidates will address themselves to the same problem, do, however, deserve careful consideration.

Criticisms of the Essay Test

Difficulty of Evaluation

The most recurrent criticism of essay tests, and the one about which most has been written, concerns the unreliability of evaluating essay answers. If a test is to be worth while as a measuring instrument, it must measure what it purports to measure consistently and dependably. A test which elicits responses which are not evaluated consistently by competent and well-trained readers cannot be of value as a measuring instrument. Consider an actual case from the experience of the College Entrance Examination Board (8). A test in English composition, designed to measure the extent to which the candidates could express themselves with clarity and accuracy, contained a question requiring the candidate to write

a paper of some 350 words on one topic selected from seven which were given. The grade on the question was intended to be an index of the candidate's ability to handle the English language, an important matter. Answers to this question were written by 6,834 secondary school seniors from a variety of backgrounds and from schools located in almost every section of the country. Readers drawn from secondary schools and colleges were invited, because of their special competence in the field of English, to New York to read the papers under supervised conditions. Many of the readers were experienced in this work. After discussion of how to read the papers, development of a scale of values, and some practice sessions, the reading got under way. Each paper was read twice independently; that is, a second reader graded the paper without reference to or knowledge of the grade assigned by the first reader. Each paper was assigned by the reader to one of eight possible quality groups according to the reader's judgment of the worth of the paper. The second reader agreed with the grade assigned by the first reader in only 41 percent of the cases. The correlation between the grades given by the first and second readers was .55—about the same as the relationship between height and weight. Of this group of 6,834 papers, 74 were graded as bad failures on the first reading, a category used only for papers which were so poorly written as to deserve a classification worse than mere failure. This grade had a special meaning to the readers: the papers so rated were said to indicate clearly that the examinee could not possibly handle freshman English at any college. Yet of these 74 very low papers, many of which were almost unintelligible or had so little written that no grade was possible, only 26 were classified as bad failures by the second readers, and 10 of them were classified as passing or better. The question then naturally arises whether one particular paper is so badly written that one can fairly say that its writer cannot write with clarity and accuracy, as the first reader judged, or whether it indicates work above passing, as the second reader judged. As handled by this group of readers, the grades on this question furnished little evidence with regard to the writing ability of the candidate. Obviously, unless essays can be read with consistency, the results cannot possibly indicate the ability in English of the examinee.

The literature on testing contains many other studies showing that the typical essay test as typically handled, whether by the classroom teacher or by "experts" as in the case previously described, is not reliably graded and, therefore, cannot stand alone as a good measuring instrument. When the papers are evaluated by different teachers who, though in the same school, do not know the pupils who wrote the tests, grades generally varying widely

from those given by the original readers are likely to be assigned. If the same papers are independently read by teachers of different schools, a still wider variation in grades is likely to occur. Some differences should be expected, for just as the passing grade varies from 50 in some school systems to 80 in others, and the interpretation of what constitutes a passing grade will vary even within the same school, so the rules and conventions of writing, as well as emphases on phases of the subject matter, are different from school to school, and in some subjects from teacher to teacher. Part of the discrepancy may also be caused by the fact that frequently teachers who know the pupils grade the papers in part on the basis of judgments previously reached about the pupils rather than independently on the basis of the quality of the paper. Pet theories and idiosyncrasies of the readers almost inevitably affect the evaluation of the essay question. The matter is further complicated by differing conceptions of passing and failing and different interpretations of marking systems. But studies reported in the literature on the unreliability of the conventional reading of essay answers go even further. Usually even the same teacher is inconsistent in the grades he assigns the same papers when an interval of a week or even less elapses between his two independent readings. The variation is usually more marked if extended answers are involved. Such questions as "Discuss the place of religion in the modern world" or "What do you think of the increasing concentration of power in the Federal government?" when used as thought questions (where direct answers have not been supplied in the textbook or class lectures and discussions) result in answers presenting the greatest difficulty to readers, and are seldom read with sufficient reliability to justify their use except as learning exercises and motivating devices.

In answering essay questions requiring extended answers, the examinee is given an opportunity to exercise his skill and his higher-order mental abilities in a more natural, integrated way and to an extent which may not be possible with objective items, but the answers do not typically reveal these skills and abilities in a way which the readers can recognize and assess. In preparing an extended answer to an essay question, the examinee may use his powers of organization, his judgment, his knowledge of the field, his originality, his ability to select pertinent facts and to marshal his arguments, and so forth, but unless readers can be trained to detect these elements (which, of course, implies that they must be revealed in sufficient quantity to be detected) and to evaluate them consistently and validly, the examinee's answer is of little or no value as a means of measuring these higher-order abilities.

Reliable reading of essay papers is possible, however, where the questions are carefully framed with the problems of evaluation in mind and where

readers are trained in the techniques of consistent reading. In the early 1930's, Sims (*5, 6, 7*) reported a series of studies which indicated that essay questions, of the discussion type as well as short-answer, could be read with a rather high degree of reliability, even by persons not too skilled, if certain simple rules were followed. Extensive and costly studies conducted by the College Entrance Examination Board, and reported in their *Annual Reports,* give ample proof that reliable reading can be obtained by methods described later in this chapter. Some unpublished studies of essay questions requiring extended answers used by the Department of State show high intercorrelation among the grades on different essay questions read by readers in different universities who were not subjected to special training or instruction. The evaluation of essay questions can be handled with reliability, although the process is apt to be difficult, time-consuming, and expensive. Such a costly expenditure of both money and talent can be justified only where the questions are carefully designed to measure important objectives not validly measured by the newer type of test questions.

Many workers in the field of testing have concluded (without sufficient evidence to warrant it) that there is no solution to the unreliability of grading essay questions and that, therefore, other means of testing must be developed. The consistency of reading objective papers, they point out, is practically perfect. On the other hand, some teachers wonder whether the use of objective tests may not in some cases obtain reader reliability at the expense of validity. Do the objective questions measure all of the educational goals of any great value? No one can seriously defend neglecting important outcomes merely because others can be measured consistently, cheaply, and efficiently. Yet a question cannot be justified as a measuring instrument merely because it gives the examinee an opportunity to exercise important abilities which cannot be assessed from the product. Consistency in reading papers is not an end in itself, but so long as tests are used to measure, and so long as the resulting scores are used as an important basis for action, ways and means must be developed to secure reliable reading of examination papers. Within certain limits this problem can be solved and has been solved.

Cost of Reading

While reliable reading can be obtained under certain conditions, the cost of reading essay questions is apt to be high, and in general the better the reading, the more costly it is. Probably the College Entrance Examination Board has had one of the most carefully controlled reading sessions for essay papers, and their cost figures for reading are, therefore, especially significant. In 1941, when costs were low compared with those of 1949, it cost them $2.42 per candidate to read an English paper for which the candi-

date was allowed three hours. On the average the group of 94 readers read 1.1 papers per hour per reader, although no reader read more than a section of any paper. Because of the large number of papers—5,715—mass operating economies were possible. The papers in American history of similar length cost $1.33 per paper to read. Expenditures of this magnitude are not possible in most situations of mass testing. Granted that other methods might reduce the cost, the problem of time and money becomes important in large-scale measurement projects and, to a lesser extent, in large lecture courses in the colleges of today. It is less significant in a class of normal size handled by a single teacher, although even here the overworked teacher or the one seeking more free time for research or leisure gladly turns to objective tests, especially those prepared by others, to free him from the drudgery involved in reading essay questions. The accurate evaluation of a well-developed essay question is a long and difficult job and one which, properly done, requires intelligence, diligence, and consistency. The expense in time and money can be justified only to the extent that essay items are developed to measure reliably important objectives which cannot otherwise be measured.

Sampling

If an achievement test is to be of value as a measuring instrument, it must secure a sample of student behavior which is indicative of the student's attainment of course objectives deemed important, such as skills, techniques, methods of reasoning, and knowledge of the field. The essay test has been criticized on the ground that the sample of student performance it usually obtains is too restricted to be dependably representative. This criticism deserves consideration.

The essay test contains a smaller number of independent units than an objective test made for the same length of examination time. In an essay test, the student writes more, but the independent judgments made by the reader are fewer. In an objective test of 150 five-choice best-answer items, there are 150 independent responses by the pupil, each of which can be separately evaluated. In an essay test of ten questions, fewer independent judgments can be made of the student, although the judgments are based on a larger amount of student work. In the objective test the student reacts quickly to a large number of highly restricted situations where the response form is very limited and usually fixed, whereas in the essay test—particularly of the type calling for extended answers—the student expresses himself at length on a few, perhaps even only one or two, topics.

The size of the sample of student performance necessary for a dependable estimate of the attainment of an objective can be determined experimentally in each practical situation, and varies greatly with the situation. The size

of the sample needed to determine the quantity of chemical in solution, for example, is small; a different and more extended sample is required to determine the contents of a carload of miscellaneous express packages ranging from bed springs to strawberry jam. The size of sample needed to determine the extent of a student's vocabulary is relatively small both in time required of the student and in number of objective items, and the sample is easily obtained. The size of sample of writing necessary to determine the extent to which a student can express himself with clarity and accuracy is certainly larger, and such a sample is difficult both to secure and to evaluate. Almost all of the important abilities for whose measurement the essay test has been felt to be especially suited are abilities which would seem to require extensive samples. Furthermore, the size of the sample required depends upon the correlation between independent units of sampling. When essay questions are used, the lower the reliability of reading, the larger the sample required.

How can one determine whether or not a test does adequately sample student performance? After the objective to be measured has been defined in some way and a test developed which is believed to measure the attainment of this objective, one can check the results of this test with other evaluations of the attainments of the objective. Frequently teachers' marks are used as the criterion. Teachers' marks are not usually satisfactory, however, because of the many extraneous factors which enter into them, such as handing in papers on time, classroom behavior, and so forth. They should never be used as a criterion for evaluating the adequacy of the sampling in examinations prepared and read by the same teachers who assign the grades. In grading essay examinations, if the pupil is known, there is a tendency for teachers to read into the results judgments already reached concerning the ability of the pupil. In situations where many tests can be used, the results of the new test can be checked against the cumulative results of all the other tests. If the new test results check closely, the sample is adequate. If an existing test is known to measure the ability, a new test can be correlated with it; if the relationship is high, the new test does an adequate job of sampling.

As has been said, the essay test must be reliably read. More than this, it must be reliable in the sense that the score on the test is indicative of what the candidate will do on another similar but different test. Unless it is reliable, it does not secure an adequate sample. It is not possible to develop a reliable test of an unreliable quality or skill, and in many cases it may well be that the essay test is used in an attempt to measure abilities or skills which are unstable or even nonexistent. For example, a person may be said to be creative, but it is possible that creativeness is an unstable trait. One

cannot necessarily say that because a person is unable to be creative at a certain time, under the terms laid down in an examination, he is therefore not a creative person. A poet of some established reputation may not be able to create a poem on the examination day at the specified hour and under the conditions of an examination. It cannot correctly be concluded that he cannot therefore write poetry. Many of the abilities for which essay tests are said to be especially suited (which will be discussed later) may be abilities which, if they exist as general abilities, are unstable or erratic, and hence relatively difficult to measure. It may well be that the sampling in the essay questions should be extensive in the sense of spreading over a considerable period of time and a considerable variety of situations if we are to make even a first approximation at measuring these higher-order abilities. The ability to organize, to think clearly, to write clearly, to use learning creatively—these are intellectual abilities generally praised, but probably no examination of two or three hours in length can ever be developed to measure directly these so-called abilities. If they exist at all, as generalized abilities, they do so only under rather special and narrow circumstances. The essay test should not be condemned because it does not weigh the imponderable or measure that which does not exist or which at the particular time does not choose to display itself.

If an essay test adequately samples student performance, it will agree with another test of the same characteristic. The more unstable the trait being measured, the longer the test required to secure an adequate sample. Some of the abilities for which essay tests are claimed to be suitable measures may require a test of almost infinite length. Conversely, a small representative sample may adequately measure a relatively homogeneous field.

Practically speaking, and considering measures of achievement in the usual classroom subjects, no measures have yet been shown to be highly reliable unless they contain a number of independent responses. For example, if the essay test contains only a few questions, each calling for a long answer, the examinee who happens to misunderstand one question will be penalized disproportionately to his real knowledge of the subject. No such serious damage is done if for similar reasons he misses a few of the 150 items which could be set in an objective test of the same duration. Objective tests are generally more reliable, hour for hour, than essay tests. The essay test may be suited to probing deeply, but no essay test of one item has yet been shown to yield a stable grade, no matter how deeply it probed. Therefore, until better techniques of handling the extended response have been developed, the teacher might well use essay questions of more limited response and include more questions. Essay questions requiring only brief answers can be devised which measure significant achieve-

ments not easily directly measured by objective tests. Another way of overcoming the limited sampling characteristic of the essay test might be to use a series of essay quizzes throughout the term and accumulate the results for a final grade which will be reliable.

Another issue in essay examining which has implications for sampling is that of the use of optional questions. A test situation is one in which the candidate is asked to show his competence by responding to certain definite questions. The questions asked are intended to sample the field and thus provide a sound basis for determining the extent to which a student has mastered the course. In all cases of scientific sampling a definite procedure is prescribed and followed. Once a sample is taken, it is not discarded without being recorded. When the quality of a carload of wheat is to be determined by sampling the wheat, the buyer does not select fifteen samples and then ask the owner which ten he wishes to have used. Such a procedure would be patently absurd. It is equally absurd in a spelling test to allow the candidate to select the words he wishes to spell. Yet with essay tests such a procedure is common. The candidate is presented with, say, fifteen questions and told to select any ten. The theory supporting the use of optional questions is that the measurement is not for possession of knowledge or of facts, but of an ability or skill which is independent of the particular question used. For example, if one wishes to measure the ability of the examinee to write concisely and accurately, content is theoretically of no importance and the ability can be shown equally well by writing on any of a variety of subjects. Therefore, the examinee should be given the opportunity to select the subject on which he thinks he can do best. Of course, his judgment may be wrong—he might be able to do better on another topic than the one he selected. If the dubious assumption is made that when one can write concisely and accurately on one subject, he can do equally well on another subject, then no optional questions are necessary. If one assumes that the examinee has the ability being measured if he can write concisely and accurately on only the one topic he selects, but not on another topic, then the significance of the possession of this ability is so difficult to interpret that one may question the use of the results.

An illustration from history examinations where the use of optional questions is general may demonstrate other considerations. The examiner argues that he expects the examination to determine the ability of the examinee to marshal data, organize them, and present them clearly, rather than to measure knowledge of specific facts or any special topic. He prefers, in fact, to allow the examinee to elect a topic on which he knows the most. Therefore, he uses optional questions. A single question, in such a case, can usually be devised to cover this situation. However, the measurement

of abilities to organize, to write clearly, etc., cannot at this time be measured independently of the topic in which the writing centers. It is, therefore, best perhaps to recognize this fact and to have all examinees address themselves to the same basic problem.

One other problem is intensified by the use of optional questions. It gives the examinee greater opportunity to use material prepared by others which he has memorized for reproduction in the examination. It is sometimes not too difficult for him to calculate the probability of certain topics being included among the optional questions. With the aid of more able and experienced persons he can develop an answer which will please the reader. In such a case the answer obviously does not represent the ability of the examinee, and the examination question is therefore not a valid one.

It is possible that the use of optional questions, which complicate the sampling problem in an unknown fashion, continues in favor primarily because the paper including optional questions appears to the teacher as a more complete representation of the course; he disregards the fact that the questions actually answered by any one examinee represent a limited and perhaps distorted sample of what the student knows of the course.

No experimental evidence has been published to show that skills and abilities can be adequately sampled by the use of optional questions; on the other hand, several studies have shown that optional questions complicate measurement and introduce factors of judgment which are extraneous to the ability being measured. For sound sampling, it is recommended that optional questions be avoided and that all examinees be asked to run the same race.

Technical Criticisms

The essay question has been criticized for faults which are common to essay questions as they are used, but which are not inevitably characteristic of the essay form. For example, readers of essay questions frequently read papers in terms of a passing grade. If ten essay questions are to be answered on a paper on American history in a school where 70 is the lowest passing grade, then each question is counted 10, and the reader thinks of the answer graded 7 as passing, that graded 8 as average, and that graded 9 or 10 as superior, while 6 or below is failing. Such a procedure reduces the reliability of the measurement through increasing the subjectivity resulting from varying grading systems and through restricting the range of the marks. Marks so arrived at obviously could not be compared with marks from a school where a different lowest passing grade was used. The restricted range of the marks so arrived at is also obvious. The poorest papers will probably receive a total score around 50 and the very best of around 85 or 90. The

mean score will be 75 and the standard deviation 5 or less. A wide distribution of scores is desirable because a wider range is usually more descriptive of the actual distribution of the ability being measured. Essay questions should be read on a point scale without reference to the passing grade, with the mark accepted as a "raw score," and the qualitative interpretations introduced after the reading is completed. In other words, points may be assigned for subject matter, style, organization, and so on, so that each answer is graded in various categories without reference to a passing or failing grade. The total scores can be interpreted qualitatively after the reading has been completed. While, in general, the essay question requiring, say, fifteen minutes cannot be read so as to yield a score range as great as that obtained by fixed-response items for the same time interval, it can be read so as to give a greater score range than has been common.

Finally, another fault associated with the essay question that is prepared and read by someone other than the classroom teacher, which affects the dependability of the sampling, is the confusion on the part of the examinee in knowing what is wanted. Students in the schools of England are trained to answer at length such questions as "Discuss the literature of the eighteenth century." Our students, not so trained, must usually be given rather definite instructions as to what is expected. In the classroom situation, the pupils usually master very quickly the requirements of an individual teacher and will answer his questions accordingly. When confronted with tests prepared and to be read by others, however, the pupils are likely to be at a loss to know what length as well as kind of response is required. It is, therefore, desirable to give explicit instructions as to both time and length of response for each question.

Potential Measurement Values of the Essay Question

Freedom of Response

A fundamental characteristic of the essay question is the freedom of response which it allows. The candidate is not given fixed responses from which to select a single word or phrase which is the correct answer to a question, but he is instead required to select from his own background and knowledge the pertinent information, to organize the answer, and actually to express it in his own words. In the practical test situation, the examiner should devise questions that define the limits within which this freedom should operate. The essay test can be a test of how the candidate approaches a problem, what information he thinks important, what conclusions he can reach independently. It can do all of these things and still be directed to a specific area of a definite field.

A few elementary and simple illustrations may show how in even a very short-answer essay question the candidate has freedom. Consider the question, "Colors are seen in a thin film of oil. Why?" Here the examinee is told of a simple observation which his own experience confirms. He is asked to explain the reason for the phenomenon. From his knowledge, he must select pertinent information, decide how best to express it, and write the answer. From a subjective analysis, the process the candidate goes through in answering this very short question appears to be basically different from that he will use in answering a similar question in a fixed-answer form:

> Colors in a thin film of oil are caused: (1) wholly by reflection; (2) wholly by refraction; (3) by interference; (4) wholly by absorption; (5) by radiation

Another type of question makes more obvious the possible difference in the thought process used to answer a fixed-answer and a free-response question on one part of mathematics:

> Two unknown instruments, w and s, are observed to vary, and the following simultaneous readings on the two instruments are noted throughout the day: 2 and 18, -1 and 3, 4 and 48, 0 and 4, 1 and 9, -2 and 6, 3 and 31. What relationship exists between the two readings, and how is it determined?

In this case the candidate must apply appropriate rules to reach a result. He must organize and present his reasoning leading to a specific conclusion. If, instead of the essay question, five equations were given and the candidate asked to select the correct one, a substitution of numbers could be used and the reasoning required by the essay question omitted entirely. Of course, the candidate could use the more elaborate reasoning process in the objective test item also, but he may not need to, and whether he does or not is not shown by his answer.

The essay question should usually present a definite problem. The examinee must be directed into a specific situation and then allowed to react freely within the limitation imposed. By making clear the nature of the question, one does not do away with the freedom of response. To the general question "What do you know about the periodic table?" the typical student of elementary chemistry would not know what to answer; and probably the examiner who would ask it would not know what he wanted. Such questions suggest no effort on the part of the examiner to identify the outcomes of learning concerning which he wants information. The response to such a question cannot be evaluated by present techniques. Dozens of approaches might legitimately be made; and what each one signifies, no reader can know. This type of freedom—license, not liberty—should be

avoided. It leads only to confusion, not to measurement. A fundamental characteristic of the essay question is the freedom of response it allows, yet the degree of freedom allowed should be appropriate to the outcomes one is seeking to measure, and freedom is possible in answering definite and limited questions.

Direct Approach to Important Goals

One of the chief claims for the essay test is that it measures important outcomes of education which are not otherwise measured. This claim is generally accepted without proof, evidence, or supporting logic, and is usually taken to mean the larger goals of all education. Sims, in the article referred to earlier in this chapter, recognizes this fact. He states, ". . . the value of the essay for testing ability to organize, relate, and 'weigh' materials learned has been long appreciated. These, however, are but a few of a number of higher mental processes for which the test seems well adapted. To name others, where the concern is over the extent to which particular learnings are integrated with previous learning, over the 'distance' that particular learning can be transferred, or with the ingenuity of response (the ability to use learning creatively), the essay examination seems called for" (*4*, pages 19–20). But, he labels such claims as nothing more than "reasonable hypotheses worthy of being tested."

Let us consider this matter further. Take, for example, "the ability to organize." The extended essay answer must be organized, and, therefore, it is commonly assumed that an essay answer will furnish evidence of the ability to organize. The ability to organize a piece of writing is, however, an involved and complicated ability. Perhaps there is no such thing as a general ability to organize. One may organize one paper well and another poorly. How many examples of organization are needed in order to judge the ability of the writer to organize? Then there is difficulty in recognizing "good" and "poor" organization. Under test conditions by one method of evaluation, at least, the writer found that instructors of English composition who were teaching problems and methods of organization were unable to agree on the quality of the organization of student papers, whether an outline or the actual written paper was being judged (*9*). The relationship found between the scores on an outline and on the organization of a paper on a different subject was .23, which is of course in part the result of low reliability of reading the papers. No evidence has been presented which shows that a particular essay test is a dependable measure of the ability to organize, or that judges can agree very closely on the quality of organization shown in a particular set of students' papers prepared under test conditions. These facts do not deny that the essay question may be suited to

measure how well a student can organize, if such an ability exists and is stable enough to make measurement of value. It is unfortunate that little research has been done relating to such claims for the essay test. It may be desirable to analyze the components of such concepts as the ability to organize. Of what does the ability consist? Possibly its elements can be better measured separately.

Similarly, none of the other broad higher-order abilities for which the essay test has long been considered especially suited has as yet even been established as subject to dependable measurement, let alone measurable by essay questions. Consider the ability to use learning creatively. Such an ability is one which educators would favor. Instruction should be directed toward its development. But first, what do we mean by using learning creatively? How can a pupil show that he can use his learning creatively? Is it a stable ability, always on tap? If one used learning creatively in repairing a broken chain on a bicycle, can he do equally well in solving a novel problem in mathematics, in working out a practical solution to a local political situation, and in developing principles of logic? What elements are involved in using learning creatively? Such questions can be answered only as we develop adequate hypotheses and techniques of measurement. The free-response type of test should be used in experimentation for such purposes, but not as a measurement of such hoped-for abilities in the usual test situation. Often when careful analysis is made of these higher-order abilities, and operational definitions developed, essay questions of a limited variety or even objective tests can be developed to measure aspects of them. In the present state of our knowledge, however, such uses of essay tests should be limited to experimental situations.

On the other hand, worth-while learning is not restricted to the simple recall of factual material. The suggestion here is that restraint be used in attempting—in practical test situations, not in research—to measure the more high-sounding outcomes which all favor but few can define, or, having defined, can recognize. Essay questions aimed at less grandiose outcomes will probably—by requiring the candidate to compare, to contrast, to present evidence, to solve new problems, to approach the subject in a new way—measure, directly or indirectly, some higher abilities and contribute to knowledge about the candidate. The essay question suffers because many teachers expect it to measure attainment of objectives that probably cannot be measured adequately in the typical test situation and possibly cannot be measured at all.

The essay test by its very name suggests a general and highly skilled treatment which thousands of candidates who answer essay questions are not

able to handle. A case of testing in English composition illustrates this point. In England in 1931 some 67,000 youngsters were examined on their ability to write English. They were asked, much as thousands of American school children are asked, to write an essay on one of the following topics: *Gossip, Gardening, Trees, Gambling, The Good Old Times, Pilgrimages, Changing One's Mind*. Sir Philip Hartog, in discussing this examination in *Essays on Examinations,* says that no schools would object to essays of this kind because the subjects are the kind set at the schools. Yet sixteen-year-old youngsters were asked to write anything about one of these subjects for anybody. The school boy is to be learned, witty, charming, spontaneous, and show himself the equal of Addison, Johnson, Goldsmith, or Lamb. It is an inappropriate task and a gross but common misuse of the essay question. The results are doomed to be worthless as measures of any ability of importance or dependability. Superficial consideration may suggest that this type of question measures higher-order abilities—organization, clarity of thought, wit, ingenuity, originality—but actually it has yet to be shown to measure anything of any significance whatsoever.

The essay question is not benefited by the misuse to which it is frequently put nor by the impassioned defense made for it by those who seek to measure these large, complex, but worthy goals which are generally felt to be important but which have not been defined or understood. The essay question may well have values for such purposes which are unique, but this fact can be established only by careful experimentation, which has in large part yet to be done. In the meantime, it would seem advisable to direct the teachers' use of essay questions toward immediate outcomes of importance which objective tests are not believed to measure satisfactorily. This end can best be accomplished by first defining or describing the objectives to be measured and doing so in an operational fashion. If appropriate questions are to be designed, the objectives of the course should be clearly conceived. Then questions which aim at measuring the degree to which the examinees have achieved these objectives can be framed and tried. The problem is how to frame an actual question to measure the attainment of specific objectives of a very real course of study of existing pupils. Essay questions under proper conditions may measure understandings which are difficult to measure by fixed-response, objective questions. Exactly what any item measures depends on the training of the examinee who takes it and the way in which the results are evaluated. It is advisable to have the test item present a novel situation—but one not so completely foreign as to be incomprehensible—and call for meaningful understanding of the subject. Such a result may be achieved by asking the examinee to relate, compare,

contrast, discriminate, note limitations of data, draw inferences, detect relationships, select pertinent information, state conclusions tersely, interpret, establish sequences, etc.

Emphasis on Whole

One advantage of the essay question is that it usually directs attention to and places emphasis on a larger segment of the subject, or on an integrated total unit. The objective test typically deals with a series of small isolated units. There are advantages, particularly in the measurement of understanding of a field of study, to having the examinee consider integrated sections of that field as well as isolated facts. Such a question as "Compare the success of Grant as a general with his success as a president," whatever its faults, does force the examinee to consider Grant's career in its broader aspects. Such a question should lend itself to reliable reading. "What are the fundamental differences in government between a democracy such as the United States and the English form of government?" This question forces the examinee to consider the basic characteristics. "What are the principal conclusions which can properly be reached from the data given in the following table?" Here again the examinee must consider the entire table, determine what conclusions are both pertinent and justified, and express his conclusions in his own words.

Attitudes

The essay question is considered to be one means of measuring the attitude of the examinee toward problems and subjects considered in the class. The measurement of attitudes, the change in attitudes as a result of a certain course, the strength of attitudes, and relationship between attitudes and action, are all significant educational problems. Generally speaking, the measurement of attitudes is commonly not considered a legitimate part of achievement testing when the results are used for purposes of helping to determine a grade. The extent to which measurement of attitude should be a part of an achievement test would seem to be a matter of the objectives of the particular instruction. So far as the measurement of attitudes is concerned, objective tests can be prepared which indicate attitudes if there is no compulsion to simulate a particular position. In answering an essay question, however, especially of the type requiring an extended answer, the examinee may reveal his attitude on basic issues without intending to do so, or he may profess an attitude which is not genuine. While no experimental evidence has been presented to show that answers to essay questions are a suitable or efficient method of measuring attitudes, attitudes are undoubt-

edly revealed to some extent in answering appropriate essay questions. What is needed, of course, is experimentation to determine the extent to which attitudes are so revealed, the kinds of questions that cause the pupil to reveal them, and suitable methods for evaluating them.

Creative Ability

The essay question has been proposed as a suitable type of question to use in securing a sample of the creative powers of the examinee. However, the fixed and formal examination period is not an ideal time or place to have creative efforts made. While the essay question can ask for the creation of poetry, essays, and other original contributions, such abilities, as has been pointed out, are not usually available on demand. They are unreliable abilities and probably should not be asked for in such an examination. If, however, they are to be tested, only the essay question can do it directly.

Essay questions on scientific topics, as well as those in the humanities, may call for certain types of creative ability: the power to apply basic principles to the solution of a novel problem is in part a creative power. Still, until "creative ability" has been more precisely analyzed and defined, it will be wise to set essay questions, whether on scientific or other topics, which demand a relatively small amount of "creativeness," and to have many questions required of all the examinees.

Insight into Personality

The free essay question demanding extended answers permits the examinee to display certain characteristics, and from his writing a reader may be able to infer correctly some things about the examinee's personality. Obviously, in achievement testing the interest of the examiner would be confined to personality changes that are consistent with the objectives of the particular learning experience. Inferences about the personality of an examinee from answers to examination questions are difficult to draw, and the reliability of such conclusions should always be checked. The basis for unexpected insights into the personality of the examinee may be revealed occasionally by even the most sophisticated examinee in his answers to essay questions, but the examiner cannot depend upon such revelations, and their detection and accurate interpretation are matters of difficulty, perhaps more comparable to the interpretation of anecdotal records than to the evaluation of responses to examination questions. Nonetheless a shrewd teacher *may* learn a good deal by the careful study of the answers of his students to appriate essay questions. Such study is recommended and suggests another use of tests than the conventional function of measurement.

The Influence of the Essay Question on Teaching and Learning

Pupils adapt their learning to meet the requirements of the test situation. This adaptation is to be expected because the tests are the most tangible cues and the most potent single influence in determining the goals of study. The typical pupil is anxious to do well in school work. Parents stress success in school, and the usual incentives to excel are present. Since success in school is measured in terms of marks and since marks are based in a large part on the results of tests, the pupil soon learns that the test is the real hurdle. He therefore directs his learning in the paths which he believes will lead to high grades on the test, and if there are short cuts, he will use them.

What type of study does the essay test foster? Experimental studies (for example, those by G. Meyer [*2*] and by P. W. Terry [*10*]) suggest that when students know they are to be tested by essay questions their study methods are different from those used when they prepare for the usual objective test. The essay question serves as a worthy goal. It causes the pupils to consider the important ideas, to outline and compare, to develop applications and illustrative material, to place the topic in its field. Objective tests, as commonly prepared by the classroom teacher, tend to encourage memorization of isolated facts and the cramming of minutiae.

The essay question can give the examinee an opportunity to exercise his judgment and to organize and write. It follows that in preparing for such a test the pupil directs his attention to the larger aspects of the subject, and attempts to master the skills and techniques, and to acquire the knowledge demanded. The objective test, on the other hand, consists so frequently of isolated factual information that the pupil tries to acquire as many facts as possible and sees no need to consider the broader aspects and meaning of his subject. Unless information is integrated and associated with other learning, it tends to be forgotten. Thus, it is to be expected, and some experimental evidence confirms, that material studied for essay tests is retained longer than when the preparation is for an objective test.

The value of essay questions in fostering sound study habits should be recognized and exploited. Questions should be developed with this end in view. If a question is to be a valid measuring device, it must do more than stimulate appropriate study habits, but suitable essay questions can do both.

The fact that the essay test fosters sound study habits suggests that the essay question can serve also as an incentive for both teacher and pupil. Because the essay question can deal with an integrated large unit of subject matter, because the possibilities of response are not limited by the ingenuity of the person constructing the question, but are as broad and as

complex as the examinee chooses to make them, the essay test is especially suited to serve as a goal. Certain aspects of the essay question which make it difficult to evaluate tend to make it an excellent goal for study. Even though the question elicits responses which cannot be read reliably and the question is not therefore suitable as a measuring device, it may serve to stimulate the pupils to strive for abilities and knowledge which they believe are necessary to answer the question. The teachers, likewise, may use essay questions, especially the broad type of question requiring an extended response, as an appropriate indication of the type of problem the students should be able to handle as a result of a particular course.

Thus, the essay question can serve to stimulate sound study habits, to direct attention to the more important aspects of the subject, to encourage students to think of larger units of subject matter and more natural complete units. It can do all of these things and still be a poor measuring instrument. For example, an objective vocabulary test is the best general means of predicting success in elementary English composition. It is a sound measuring instrument, but as a stimulus for appropriate study, it has almost no value —it may even have a detrimental influence. A test which requires the examinee to produce an extended piece of writing—one demanding organization, giving an opportunity for the examinee to demonstrate his style of writing, his choice of words, his broad appreciation of literature—this type of question is of excellent use to stimulate the students, to direct their study to the final large goal. It does not result in an answer which thus far can be handled with a high degree of reliability, but it does have definite values in the teaching situation.

One value of the essay test is that it requires the student to *express his ideas in writing*. This value, again, is quite independent of the worth of the question to measure skills or abilities. For teachers who are genuinely interested in the education of their pupils, this value will be of great importance. Granted that the objective test can be given easily and speedily and can be scored cheaply and accurately, nevertheless, it can never compel the student to think out for himself what is to be said on a given topic, how it is to be said, and to perform the actual writing. Skill in these abilities is so important in education and in life that procedures tending to increase this skill should be cherished and developed at whatever cost. Not a few teachers firmly believe that the spread of objective testing has been perceptibly detrimental, if for no other reason than that pupils who know they are to be tested only objectively have little incentive for writing. The essay test provides the incentive for writing; it goes further by giving actual practice in the art of writing. The overworked grade school teacher whose success will be

measured, if at all, by the scores obtained by his pupils on objective tests involving no writing can hardly be blamed if he steadily reduces the amount of written work in his class. College teachers are finding that more and more freshmen, even those with high aptitude test scores and brilliant school records, are unable to express themselves in writing with reasonable accuracy and clarity. It is unfortunate, for this reason, that large testing agencies are eliminating all forms of the essay test.

Written and oral expression is important in the lives of most adults. With modern methods of instruction properly utilizing visual aids, talking pictures, illustrated texts, phonograph records, the radio, and now even television, with new-type tests requiring only short marks by special graphite pencils, automatically graded by electronic machines; with overworked and underpaid teachers unable to find the time and unwilling to use the energy required to decipher the typical handwriting of pupils, the pupils are doing less and less writing. There is no necessity, or even opportunity, for it in many of the larger classes today. Thus, one value of the essay test is that it gives the pupil an opportunity from time to time to write a complete sentence or two, to express his thoughts in his own words, and it holds up before him an important goal—the ability to express himself unaided by several prepared responses from which he can select the best, the most cogent, or the clearest.

One would scarcely call the examination room the best place in which to foster good writing. Examinees under varying degrees of nervous strain attempt in essay tests to produce in a given limited time a finished product. One would imagine few situations less conducive to brilliant or even sound prose. Experienced readers of essay tests, however, bear witness that not infrequently, under such heat and pressure, able and well-trained students have produced remarkable results. What is more important for the whole group of students is that, if examinations do not demand writing, the pupils will question the need for writing in the classroom, and teachers will in many instances also welcome a relief from reading the handwritten essays. It is a sad commentary to make, but in the mass education of today the opportunity to write which the essay question provides must be seized upon —despite its drawbacks—as one of the few chances the pupils are having to write.

Suggestions for Improving Essay Questions

In spite of the increasing use of objective tests and of the unsolved problems connected with essay tests, essay questions continue to play an important part in the measurement of attainment—because they attempt to measure abilities which cannot be, or have not been, adequately measured by

objective tests, because their use has certain important effects on education, because they are traditional, and because they are or seem easy to prepare. If they are to be valuable as measuring instruments, however, teachers and test technicians must devote care and skill to framing questions and to developing dependable methods of evaluating the answers.

Consider the Nature of the Population To Be Measured

In setting an essay examination, the size and homogeneity of the population to be tested is an important consideration. The problem of the classroom teacher who knows his group of from ten to fifty pupils is quite different from that of a committee appointed to build the final examination in a large course taught in a great many sections by different instructors, or from that of the College Board examiners in English, for example, who set a test which will be written by thousands of students from schools all over the United States. The question "Discuss Shakespeare's idea of tragedy" set as a question for thousands of students at the college entrance level who come from different schools and have read different plays of Shakespeare under teachers of widely different kinds of training and ability, will produce answers of such extremely varying kinds that they will be incommensurable. That same question, set for a group of students in the same college who had read the same plays under the same teacher, and who had been trained in considering the nature of tragedy in each play, might produce measurable and significant results. The limits to the answers would have been set by the training common to all, rather than by the question itself. The background and training of the group to be tested must always be considered if suitable questions are to be prepared. A question can be quite satisfactory from every point of view for the pupils of one instructor in one class and yet be unsatisfactory when applied to a large group of students from many schools covering a wide area. This fact can hardly be overemphasized.

Define the Objectives in a Meaningful Fashion

Before any achievement test questions are prepared, a careful consideration of the important objectives of the course is essential. The testing should be aimed at the important objectives. It should sample the important skills, knowledge, abilities, or attitudes the course has aimed to develop, and do so as thoroughly as possible within the time allowed. One cannot determine the outcomes which a particular question really measures unless one knows what the student has been taught and how. Many questions, including those which appear to be thought questions, require merely that the student give back the answers as they have been taught him. Possession of

basic factual knowledge is important in many courses, but where it is important, it can usually be asked for in a simple direct fashion. Since wide sampling of factual learning is usually essential, it would seem generally more economical to measure such learning by objective tests.

The problems of defining the important objectives of a course in some meaningful or operational fashion are many. It is not possible even to review them here, but obviously a clear identification of these objectives is necessary if suitable essay tests—or any other—are to be developed to measure this attainment.

Determine the Amount of Freedom of Response To Be Allowed

After the nature of the group to be tested has been considered, and the objectives to be measured have been defined, the next problem concerns the freedom of response to be allowed. The question of the freedom of the response is a serious one which must be decided before the question is prepared. Since a fundamental characteristic of the essay question is the freedom it allows the examinee, one might argue, therefore, that the greater the freedom, the better the question. One might go to the length of posing for English composition, for example, a single question, or rather directive, "Write," and permit the candidate to write anything about any topic for any purpose within the three hours allowed. For some purposes such a test might be appropriate, but measuring the outcomes of the conventional course would not seem to be among these purposes. Such a test, if it can be called a test, is not a satisfactory measuring instrument of the attainment in English composition, for example. A test situation by its very nature is restrictive. It is designed to see what the candidates can do under specified limited conditions. Broad questions of the following types are almost equally unsatisfactory as subject-matter tests: "Discuss democracy," "Write an essay on Newton's laws," "Describe a book you have read," "What does symphony music mean to you?", "What do you think of Freud?", "Tell what you know about Africa," "Tell what the study of modern European history has meant to you." These questions, it is true, give maximum freedom. Indeed, the freedom is such that no techniques are now available to evaluate the resulting answers on any valid basis. Perhaps the answers to such questions contain revealing information about the attainment and personality development of the candidate, but until techniques have been evolved to reduce this information to usable terms, such questions should be left to the research workers. More success will be derived from questions which are much more restrictive but which still give the candidate an ample opportunity

to collect his ideas, organize them, and express the result in his own words, even if the final product consists of only a few sentences. Such questions can be basically different from objective test questions. It should be remembered that there is no evidence that the essential characteristics of an essay question cannot be found in an item requiring an answer of a few sentences; nor is there any evidence that the extended, broad essay question, such as "Discuss Shakespeare," is of real value in determining the degree of attainment of the usual objectives of the school or college courses. The problem in setting an essay question is so to word it that all of the candidates demonstrate under restrictive conditions their respective achievements in important knowledge and ability. And the ability demonstrated can be of the highest intellectual order.

Essay questions which can be read with some consistency by the methods commonly used are almost always restrictive in the sense that they are questions so worded that all candidates will interpret in the same way the task to be done. Some indication is usually desirable concerning the length of the answer expected. An example of a restricted essay question, but one which still allows an extended answer is: "Explain the principal doctrines and practices of mercantilism and show how they affected international relations in the eighteenth century." The restriction may be even greater, as in the following questions:

> Answer briefly each of the following questions. The essential points in each case can be covered in a few sentences. Be as definite and concise as possible.
>
> a. What is meant by the statement that France, before 1789, was centralized but not unified?
>
> b. Why did the Dreyfus affair become an issue of national significance?
>
> c. Who said the following and to what end? "If happiness is present, we have everything, and when it is absent, we do everything with a view to possessing it." [*11.*]

The answer to this last question is in part objective and in part essay. It might be answered, but not without careful thought, somewhat as follows:

> Epicurus. It is a formulation of his well-known hedonism and an indication of the way intelligence facilitates the good life.

Such essay questions requiring relatively short answers give the candidate freedom to select from his background of knowledge pertinent information bearing on a particular problem, to decide how best to present this information or reasoning, and to compose and write down the answer. There are several advantages in forcing the candidate to address himself to a particular narrow problem. It is a real situation. The task is defined. The examinee knows what is required, and is not put into the situation

where he must guess what some particular examiner had in mind. All examination questions have some restrictions, even the broadest projective type of question; the questions requiring a brief response merely restrict the field to a somewhat greater extent, while still allowing the candidate ample freedom. Such questions have all the fundamental characteristics of the essay question, and they can be read with reliability. They should be used more widely.

Check on Technical Considerations

Before an essay question is put into final form, the examiner should consider a number of points. Has the question been phrased so that its purpose and limits are clear? For the larger, more heterogeneous groups even more explicit definition is necessary. Years ago, one of the possible theme titles on a College Board English paper was "The Vanishing Horse." The examiners had in mind an essay of a more or less economic kind on the disappearance of horses with the coming of the motorcar. But the readers found that not a few students accepted the title as a challenge to create a fairy story, and one theme with an improved title, "Rudolph, the Vanishing Horse," will never be forgotten by those fortunate enough to peruse it. Needless to say, it was found to be almost impossible to grade on the same scale used for the sober accounts of the disappearance of quadrupeds as prime movers.

Can the examinees writing the answers be expected to have adequate material with which to answer the question and can they understand the wording of the question? There is no advantage in confusion or in using words which are not understood. One question in an examination in English composition, for example, intended for a large and heterogeneous group, dealt with the interpretation of the phrase "the dog in the manger," which it was supposed would be familiar to all. It was discovered, however, that many pupils from urban communities had not the faintest notion what a manger was, and the question was, therefore, not a suitable one for all the group being tested.

Can an adequate answer be developed and written in the time allowed? Too frequently the examinee is rushed at an almost impossible rate in essay questions. It is often wise to indicate the time limits for each essay question and the time limit should be set in full consideration of the wide variation in the speed of writing found among pupils of the level being measured. For an important test it is best to try the questions out on a similar group to see what takes place under test conditions. Time should be allowed for revision of the answer, and thus all examinees are given some leeway.

Analyze the Items

The problem of how the answers are to be read should be considered before the test is given, and wherever possible considered in the light of some pretest answers which can be used to try out reading techniques. Incidentally, this procedure frequently results in a radical revision of the wording of the question. In the essay question the pretesting and item analysis must be based upon the answers *as they are read*. Different results will be obtained from an analysis of the same question answered by the same students if a different method of grading the paper is used.

Item analysis can be useful in improving essay examinations as well as objective examinations, although because of the time required and difficulties encountered, essay examinations are seldom analyzed by item response. The teacher-made test used by an instructor in his own class is seldom composed of items which have been tried and found successful. Some teachers, on the other hand, accumulate a file of essay questions and re-use them with such changes as are suggested by experience, even though no formal item analysis has been made. The test questions can be classified on the basis of the objectives which they seek to measure, and also on the basis of the difficulty of the item as determined by its actual use. The standardized objective achievement tests are usually composed of items which have been subjected to analysis, and the validity and difficulty of which are known. A test can, in this way, be assembled which will be of appropriate difficulty and directed toward the objectives thought to be important.

The item analysis techniques developed for objective items can be applied to essay questions with but few changes. Each element of an essay question which is given separate consideration should be assigned a separate point score, and these scores on elements or aspects of the response should be treated as separate items. Both difficulty and validity indices can thus be obtained. The usual criterion used in analyzing objective tests can be used. A rough scrutiny of the distribution of such item scores can itself show which parts of the essay test are discriminating among the responses and which are contributing almost nothing. What is sought is an essay item which can be reliably read to yield a wide range of test scores on a valid basis.

Some special problems are involved when several supposedly independent scores are assigned to an answer to a single essay question. The essay response may, for example, be read for accuracy of facts, method of presentation, organization, and grammatical accuracy. If each of these evaluations is independent of the others, then the scores on each judgment

should be analyzed independently. It will frequently be found either that all of the part scores are so similar that they should be combined and then analyzed or that certain of the judgments are so completely independent and erratic that they should not be counted. If in order to correct this situation a change is made in the method of reading, a check should be made to insure that the reading of other aspects of the question is not also altered. The several scores are usually not completely independent. Hence, a reanalysis of all scores is necessary.

The analysis of scores of various aspects of a single response is time-consuming and complex. However, this technique can be used to advantage in building up sound questions, answers to which can be properly read

Develop Reliable Methods of Reading

The first step toward reliable reading, as has been pointed out, is to frame questions in such a way that they call forth responses which can be consistently judged as to their quality. If we are attempting to determine whether a candidate can think clearly and accurately with certain facts, then the question must call forth a sample of thinking that is representative of the individual and one whose quality can be recognized, and consistently recognized, by competent readers. Furthermore, if the concern is with ability to reason, then the facts should probably be furnished him. Otherwise, we do not know whether lack of facts or inability to reason causes failure. In framing the questions, it is first necessary, of course, to determine and have clearly in mind what one is attempting to measure.

Once an essay question has been framed to probe mastery of certain important aspects of a course, there are two general approaches to the reading of papers. One might be called the "analytical" method, where the answers are broken down into a number of elements and each element is assigned a weight; the second, the "whole" method, where judgment is passed on the answer as a unitary whole. In using the analytical method, the objectives must be clearly defined. One of the best ways to determine the elements to be evaluated is to have the readers themselves write complete answers to the question. This step is important, and if it were taken before the essay question is assigned, drastic revision of many an essay question might result. The readers' answers are then analyzed, along with the answers of several students, in order to determine the significant characteristics of a good answer. The elements to be considered depend upon the field of study and the nature of the question. In a history paper, one might have the papers read for organization, for pertinence of factual data and logic shown, for completeness, for accuracy. Achievement of

these several objectives may be evaluated by any of the familiar rating techniques. One may in some cases even use the more objective "check-sheet" technique where the characteristic is considered as being present or absent.

This analytic approach can be illustrated in the following short essay question from an examination in secondary school physics: "Why is it that on a hot day a person feels cooler when the relative humidity is low than when it is high?" Unless the answer to this particular question has been taught, it will be a thought question requiring the application to a new situation of physical principles learned. It is a "why" question, which is a desirable type of essay question to measure understanding. It can be answered completely and correctly in a brief response. The responses can be reliably read. In determining how to read this simple question, the reader might consider first the elements that are necessary to a complete and correct answer: (1) it is common observation that on a hot day a person perspires; (2) heat is required to change water, including perspiration, into vapor; (3) the body heat is used when perspiration evaporates, making the body feel cooler; (4) high relative humidity means that the air is nearly saturated with water vapor, and cannot readily absorb more moisture; (5) thus, high relative humidity prevents the escape of body heat. These elements can be combined to give a correct and complete answer. The responses might be read by allowing one point for each of these necessary elements, and one more point for organization. Thus, a perfect answer would be scored 6. If some of the facts were omitted, but the reasoning good, a three- or four-point answer might result. All facts but not sufficient logic shown in reaching the conclusion could result in a score of 2 or 3. This particular question, when used with a large and heterogeneous group, was scored with a reliability of .85.

Experience suggests that the above described method is not reliable enough to justify its use with extended essay answers. A description of the technique of reading applied to the more extended answers in a difficult field is presented, with full illustrations, in a report published by the College Entrance Examination Board, *Report on the First Six Tests in English Composition,* by Noyes, Sale, and Stalnaker (3, page 18). On one test the examinee was given the following detailed instructions:

> Since everyone has experiences which change in some important way his opinions or attitudes, such experiences and their effects provide much material for novels, plays, and biographies.
>
> There is a wide range of experiences of this sort: a change of environment, exposure to danger, the stimulus of new friends or new ideas, a serious

illness, an encounter with misfortune. Whatever the experience may be, however, it will bring about in the person who has it, changes in his opinions, attitudes, or understanding of the world in which he lives.

Select such an experience from the life of a character you know through your reading, or from your own life, or from that of a friend. In a theme of from 400-600 words give an account of this experience and show its effects.

You will need to describe briefly the experience itself. Do not re-tell the whole plot of a story; avoid experiences which do not cause important changes. But your main problems are: to explain how the experience changed your chosen character's opinions, attitudes, or understanding, and to show by specific illustrations how these changes became evident.

Your paper will be graded on:

1. Material: your ability to choose the experience, to explain the changes it produced, and to illustrate those changes;
2. Organization of your theme as a whole and of the separate paragraphs;
3. Style, the accuracy and clarity of your writing.

The readers agreed to evaluate "material," "organization," and "style." "Material" was assigned a maximum of 7 points, but the readers were given four subheadings—experience, 1 point; character, 2; change, 2; and illustrations, 2—and the conditions necessary to obtain each point were fully described. "Organization" was allowed 8 points, proportion 2, movement of thought 2, paragraphing 2, and transitions 2. Finally "style" was assigned 8 points—spelling 2, punctuation 2, vocabulary 2, and sentence structure 2.

With extended instructions and training of a selected group of readers, it was possible to achieve in this case a reader reliability of only .58 for this long question. This is not a satisfactory reader reliability, and almost precludes the possibility that the resulting score can be of much significance. The readers themselves, after having worked arduously on improving the reliability doubted that more can be done with *this type of question.* Noyes, Sale, and Stalnaker say (*3,* page 70):

> It has been shown that a very small group of readers can reach a satisfactory level of reliability with a comparatively small number of books, but that as soon as the number of readers is increased to the point necessary to cope with a thousand or more answers, the coefficient of reliability decreases markedly. . . . The readers have tried honestly and hard to read and maintain a common standard, but in this attempt they have been only partially successful; they are not machines. It is probable that the reading of the composition test, despite its limitations, has been more reliable than the ordinary reading of themes and examinations in school and college courses. . . . Efforts to improve the reliability of reading were first directed toward reading procedures. When these efforts produced no improvement, the next

> logical step was to alter the form of the test. In September 1944 a test was set consisting of four brief topics, on each of which the examinees were asked to write a paragraph of from 75 to 125 words. In such a test, there is no chance to measure the candidate's ability to organize a considerable body of material, but this kind of measurement has proved very difficult to evaluate. The examiners felt that a candidate who could organize a good paragraph could—or could easily be taught to—organize a good theme.

The second approach to reading papers involves having the readers consider the response to the essay question as a whole, and, without analyzing it or breaking the response into its elements, use a rating technique in evaluating it. Any of the well-known rating methods have potential value for judging essay answers, although experimentation is badly needed concerning the appropriateness and efficiency of different methods for different outcomes, different types of questions, and for different teacher purposes. The reader might, for example, use the "order-of-merit" method, arranging the papers in a rank order, from the very best paper to the poorest. Since with a large number of papers this complete ranking is often not feasible, the papers may be rated by an absolute rating scale or by some variation of the "man-to-man" technique. For example, the papers may be grouped in six or eight levels of quality according to the reader's judgment of the response, and no differentiation made among the papers at each level.

In the use of this method, the readers might select papers at random until samples felt to be representative of each quality level are obtained, then have these duplicated and referred to from time to time. The reading task is then to match each new answer with the sample nearest to it in quality. Usually about six groups are all that can be differentiated satisfactorily. As many groups as can be dependably differentiated should be used. The smaller the number of groups, the more serious is a difference of one category in judging a paper. The chief problem of this method of reading is the tendency which readers seem to have to place a large proportion of the papers in the "average" category, and to avoid classifying papers in the top classification. To differentiate the levels sufficiently definitely so that the papers are well distributed among them is no easy task.

This method of judging the papers as a whole was used in grading the essays in a French examination in June 1938, which is fully described in the bulletin of the College Entrance Examination Board *Report on the French Reading Examination of June 1938,* by Jackson and Stalnaker (*1*). Sample answers at each of the six levels used are given. There are some advantages in considering the answer as a whole, and to concentrate on

a quality level may with some questions be easier for the readers than to try to go through an elaborate analytical scheme.

Whether the first or the second method of reading, or some combination of them, is employed, reliable reading of papers will result only if the questions are properly framed. When relatively restricted answers are demanded, and the readers follow careful plans and are willing to practice at the job, reliable reading can result, as has been demonstrated and reported. The person who seriously considers using either method would, however, do well to acquaint himself with the extensive literature on rating methods, the many pitfalls involved in using them, and the tested evidence concerning their best use.

Frame Essay Question to Require Specific Knowledge or Reasoning

Because of the many difficulties in understanding the total significance of an essay question without knowing the objectives of the learning, the group to be tested and the way in which it was taught and the way the answers are to be evaluated, examples of tests and test types are of limited value. Many an essay question is judged to be excellent as a measure of reasoning ability which in actuality is merely a question demanding rote memory of some reasoning taught in the classroom.

Novel or seminovel situations provide a better test of a student's understanding or ability to handle material than those discussed in the class or textbooks and repeated from memory. The problem should involve some new elements so that some reasoning is required, but not be so completely new as to be bewildering. Ingenuity in constructing such tests is an asset. One history teacher formed excellent essay questions around the headlines from old newspapers of the period being studied, and by asking students to explain the significance of selected cartoons of the day which were reproduced. Such items make good test questions and also are interesting to the examinees.

In framing questions, especially if the group to be tested is large or heterogeneous, it is wise to avoid assiduously such instructions as "discuss," "write an essay on," "tell what you know about," "what is your opinion of," and "what do you think of." Better questions can be phrased with more specific instructions. There is no reason to fear long instructions in a question. A longer question calling for a brief response is often better than a brief question calling for a long response. A request for discriminations, explanations, comparisons, contrasts, justifications, proofs, evolving of a formula will get better results than a request to discuss,

write on, and so forth. A catalogue of the possible key words in essay questions would be extensive. The following are suggestions only: *explain, relate, compare, contrast, make accurate inferences, interpret, detect relationships, show differences, discriminate, organize, select significant idea, summarize, give conclusions, generalizations, sequences, relationships, associations.*

The following questions have been successfully used:

> Explain why the propagation of apples commercially is done by grafting. [If the candidate has not been given the specific answer in class, this is a good reasoning question and can still be answered briefly.]
>
> What are the chief characteristics which serve to differentiate a corn stem from a bean stem? [Not "write an essay on a corn stem."]
>
> What steps did labor take to protect its own interests in the three decades after the Civil War? [Not "write about labor. . . ."]
>
> A coil of wire is connected to a galvanometer. When the coil is turned over, the galvanometer needle swings to one side. What is the chief conclusion you can draw from this observation? [Not "tell about electricity."]
>
> How would you prove experimentally that roots and stems of plants respond differently to the force of gravity? [A good thought question when the candidate has not been told or shown the answer.]

Where longer responses are desired, the questions can be made more effective, particularly if one is dealing with many candidates from different backgrounds, by explaining what is wanted, setting a framework for the answer. The following is an illustration of a question of this type.

> Take time to plan the answers you write on the following questions. You will be graded on (1) organization of the material, (2) intelligent use of facts to illustrate the general statements or arguments, (3) accuracy of the factual material related to the question.
>
> "The power of the president of the United States depends entirely upon the personality and policies of the man who holds office." How far is this statement true? Illustrate your arguments by reference to the administrations of two presidents.

The use of essay tests which permit a student to select any aspect of a broad subject about which he believes he knows the most and to reveal the depth of his knowledge on such a topic and the application of techniques and methods to the topic has not been developed adequately. Used in this fashion, the essay test should, it would seem, reveal special competences or depths of understanding not readily discovered by other examining means. The problem here again is how to reach a valid grade,

and essay questions are being considered in this chapter as parts of achievement tests used for the purpose of assigning grades on the basis of achievement in a course. Such a question might be:

> Give the main source of information you have obtained on any topic covered in this course [elementary sociology], the results of your study on the topic, and the nature of the research you know is now being carried on to further knowledge in this field.

Research Needs

What unique contributions to measurement of educational objectives do essay questions make? To what extent are the claims of the defenders of the essay question justified? Research is needed to answer these questions. In fact, research is needed on every phase of the essay question. It is a test form which, in spite of its widespread use, has been subjected to almost no exacting experimental study.

For example, how much freedom should be allowed the examinee in making his response to the question? Complete freedom is antithetical to a fundamental condition of examining, namely that the examinee respond to the specific questions of the examiner. In writing a completely free and extended response, the examinee may reveal certain facts about his learning, but to what extent can the examiner be sure that he has revealed his total knowledge and displayed his true abilities? Research is needed to answer such questions.

Problems of the freedom of the response and the extensiveness of the response cannot be separated from the problem of evaluating the responses, and here again research is essential. In the objective or fixed-response test, the ingenuity of the situation is a function of the ingenuity of the examiner. In the essay question the ingenuity of the examiner is equally important, but another element is added—the ability of the reader to find the hidden meaning and to detect the skills and abilities used in preparing the answer. How can the extended answer be read to reveal the significant information it possibly contains? Some techniques have been developed and have been reviewed here, but undoubtedly other techniques of greater value can be developed if based on appropriate research. For example, research on the relative value of various types of rating which have been shown to have worth in evaluating human behavior in other areas is urgently needed.

The problem of the adequacy of the sample of student behavior as revealed in the essay response is a fundamental one. The essay question is said to be especially suited to obtaining samples of the higher-order

mental processes of the examinee—his ability to organize, to show critical judgment, to synthesize, etc. To what extent are these abilities general ones, and how stable are they? What size of sample is needed before one can be sure he has a reasonable measure of them? The essay question seems to be the ideal type to probe deeply rather than sample broadly, as with the typical objective question. How can one devise questions to probe deeply so as to produce a dependable result? Essay questions should be developed which aim at the higher thought processes; then these questions should be tried, analyzed, and revised until they are in a form which elicits answers which can be reliably read and which yield scores that are reliable in the sense of corresponding closely to the scores received on similar but different tests taken at different times. These scores can then be compared with the scores on objective tests or with criterion measures based on the judgments of experts and on results of the combination of many judgments over a long period of time. Through this process essay tests could be evolved which might measure attainments and abilities not readily measured, or not measured at all, by objective tests. Rigorous, controlled experiments are needed.

Because the essay question is believed to be especially suited to measuring the broader and more significant outcomes of education, research on this question form will need to be concerned with such outcomes. A careful analysis of such objectives, the stripping of the pleasant wording down to the operational situations will be required. No easy, direct path to the answer should be expected, but there is no reason to doubt that, granted the money required—for research on the essay form will be expensive—and the time, answers to many of the quesions can be determined.

Another field of continuing study should be that of the influence of the essay question, as compared with other test forms, on the teaching and learning situation. The tested information we have here, although significant, is by no means adequate. Pupils accustomed to various types and qualities of examinations will need to be studied. The situation will not remain static; therefore, such experiments will need to be repeated.

In brief, the essay question is in need of scientific study. It has been ardently defended by those who believe in it and blithely ignored by the test technicians who are best equipped to experiment with it. It needs the analysis and study of the research worker and the psychologist interested in controlled scientific experiment. The outlining of the studies needed is a task worthy of the best minds in education and psychology, for such experiments can be of significance for the entire process of education.

Summary

The essay test has been the subject of repeated and often unfair attacks by psychologists and educationalists interested in the measurement of achievement as a science. As a result, the essay test remains largely undeveloped, although it continues to be used widely by the classroom teacher. The values claimed for it have not been generally established, yet it may well be a basic test form which, properly controlled, can measure important outcomes of learning not yet otherwise measured. It also has other potential values which have been described. It has several important and unique advantages as an educative influence. The fact that it continues to be a test form widely used by the teacher preparing his own test would alone seem to justify further development and research.

Selected References

1. Jackson, J. F., and Stalnaker, J. M. *Report on the French Reading Examination of June 1938*. Princeton, N.J.: College Entrance Examination Board, 1939. 65 pp.
2. Meyer, George. "An Experimental Study of the Old and New Types of Examination: II, Method of Study," *Journal of Educational Psychology*, **26**: 30–40, 1936.
3. Noyes, E. S.; Sale, W. M., Jr.; and Stalnaker, J. M. *Report on the First Six Tests in English Composition*. Princeton, N.J.: College Entrance Examination Board, 1945.
4. Sims, Verner Martin. "The Essay Examination Is a Projective Technique," *Educational and Psychological Measurement*, **8**: 15–31, 1948.
5. ———. "Improving the Measuring Qualities of an Essay Examination," *Journal of Educational Research*, **27**: 20–31, 1933.
6. ———. "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating," *Journal of Educational Research*, **24**: 216–23, 1931.
7. ———. "Reducing the Variability of Essay Examination Marks through Eliminating Variations in Standards of Grading," *Journal of Educational Research*, **26**: 637–47, 1933.
8. Stalnaker, John M. "Question VI, The Essay," *English Journal* (College Edition), **26**: 133–40, 1937.
9. ———. "Testing the Ability to Organize," *English Journal* (College Edition), **22**: 561–67, 1933.
10. Terry, Paul W. "How Students Review for Objective and Essay Tests," *Elementary School Journal*, **33**: 592–603, 1933.
11. University of Chicago Board of Examinations. *Manual of Examination Methods*. Second edition. Chicago: University of Chicago Bookstore, 1937.

Part Three

MEASUREMENT THEORY

14. The Fundamental Nature of Measurement

By IRVING LORGE
Teachers College, Columbia University

COLLABORATORS: Lee J. Cronbach, *University of Illinois;* Douglas E. Scates, *Queens College (New York);* Ledyard Tucker, *Educational Testing Service*

"MEASURE" IS ONE OF THE THOUSAND MOST COMMON WORDS IN PRINTED English. As is usual with words that have had a long history and wide currency, "measure" has many different meanings and applications. In a count of its occurrence in a sample of two and a half million words, "measure" occurred more than four hundred times and was used in forty different ways.

In a basic study of usage, the primary senses of "measure," as a noun, referred to all of these: the process of, the result of, the instrument for, and the units used in, measuring. Not only did "measure" mean the act or the process of determining the extent, duration, and dimensions of a thing, but it also meant the instrument by which the process is done; the units in which the instruments are graduated; and the results of the act itself. This single word does mean, then, the act of weighing, the balance in which weighing is done, or, again, the grams that are used to balance, and the numeral that expresses the result; it can mean the process of measuring, the ruler that is used, the inches in which it is graduated, and the number that gives the over-all length.

The word "measure," in addition, refers to less exact instruments, processes, and units. Among the forty meanings, it was used also to refer to any instrument used as a basis for comparison even when that comparison involved the processes of estimation or judgment. As used popularly, "measure" not only refers to procedures that have precision, but also to acts of objective estimation. "Measure" refers to the determination of the charge on the electron and also to the estimation of beauty; it refers to the determination of the weight of an automobile and also to the estimation of an individual's intelligence. "Measure" more frequently refers to acts of subjective estimate of amounts or degrees of proportion than it does to precise objective determination.

It is obvious, however, that "measure" involves even more than the lexicographer's detailed consideration of the different senses of the word. The confusion surrounding the word, may, indeed, be the consequent of assuming that the word must mean just one process, or one instrument, or one unit, or one result. When somebody uses the word correctly in one of its senses, he may still fail to communicate adequately what he intends. There are some people, unfortunately, who are even worse. They insist that the word "measure" means just what they "choose it to mean—neither more nor less."

If all scientists would agree to restrict "measurement" to a single meaning, then confusion would be minimized. Such unanimity is not an immediate prospect. Measures include the length of a table, the duration of a storm, the weight of a diamond, the resistance of an electric circuit, the achievement of a pupil, the visual response of a clam, or the expressed pain of a patient. Each of these different measures not only involves different objects but also different purposes. For instance, the length of the table may be needed to judge whether it will fit in a recess. The respective purposes for the other measures may be: making a record of the day's weather, setting a price upon a jewel, passing on the safety of a radio, evaluating the performance of a teacher, measuring the light threshold of an infrahuman organism, or selecting an anesthetic for an operation. A single sense among the variety of purposes and objects for the word "measure" is far from likely.

Most people expect that a measurement will be expressed by a quantity. In general, measurement involves the assignment of a class of numerals to a class of objects. Measurement, therefore, must consider three factors: first, it must deal with the classes of objects; second, it must deal with the classes of numerals; and third, it must deal with the rules for assigning numerals to objects. The choice of the word "numeral" in place of the more usual "number" is deliberate. For purposes of exposition, the word "numeral" refers to the symbols 1, 2, 3 . . . , whereas the word "number" refers to the meanings of the numerals. Numbers can refer to a person's social security registration, to his order in a line, to his weight on a scale. These meanings imply different relations among the numerals as symbols. Measurement is concerned with these relations.

The definition, further, refers to classes of objects. Each area of science deals with its own kind of subject matter. The physicist, for instance, deals with gases, liquids, and solids in terms of molecules, atoms, electrons, mesons, and charges. The psychologist, on the other hand, deals with behavior in terms of capacity, learning, and achievement. The subject matter or content of chemistry and of physics deals with classes of objects that are

less variable than are the classes of objects studied by the biologist or the psychologist. The differences among classes of objects necessarily involve differences in the kinds of measurement of objects. The differences, moreover, are not only between the subject matter of psychology and of physics, but also between the kinds of workers in these diverse fields. In the last analysis, as Heraclitus announced, "man is the measure of all things." It is man who studies gases and solids, who studies behavior and culture. In this sense, measurement rests on the perceptual equipment of the observer. Fundamentally, *man* makes judgments of wider, longer, heavier, and stronger, or makes estimates of quicker, smarter, and happier. Very often man extends his range of observation by machines such as microscopes and balances, or by devices such as intelligence or achievement tests. The training of scientists, to a degree, affects the kinds of observation made and the kinds of instrument used; it affects, as well, the level of precision of his observations.

The subject matter, the conceptual organization, the nature of the instruments, and the training of the scientists interact to influence the nature of the observations made. Some observations can be made directly, for instance, the measurement of lengths can be made by juxtaposing two objects, or the measurement of weight by balancing two weights, or the measurement of time by contrasting two periods of time. Other observations, however, can only be estimated from their effects. In temperature, the variations in the height of a column of alcohol or of mercury are known to be related to variations in temperature. In scholastic achievement, the variations in the quality of the performance on tests are presumed to be related to variations in the amount, or even in the quality, of what is learned. The estimate of achievement in school, just as of temperature in a room, can be observed only by its effects.

Direct observation involves perceptions of the thing or of the property of the thing itself. Indirect observation involves inferences about the thing or its properties from its effects on other things or from its own performances. The inference about a property from its effects involves either an assumption about a relation between effect and property, or a demonstration about the relation between effect and property. In the case of temperature, an established relation has been discovered between thermal dynamics and the height of a mercury column. In achievement, however, the relationship between items of information acquired and behavior has not yet been fully demonstrated.

There are many kinds of observations that could be made about things or people. For instance, a piece of window glass reflects light, transmits light, affects the path of raindrops, gets warm, etc. An individual, on the

other hand, remembers facts, pursues hobbies, develops skills, etc. So many are the effects upon things and upon persons that it is never possible to catalogue them all. To focus attention, the scientist, particularly, limits observation to a class of effects or to a class of direct observations. The distinction is between observing the thing for itself as opposed to observing it for inferences about the thing.

In scientific observations, whether direct or indirect, the conditions for observation are carefully specified in terms of time, place, and circumstance. In physics and chemistry, observations at sea level at 25° centigrade may differ markedly from observations of the same thing at 0° centigrade and in an airplane 3,500 feet above sea level. In psychology, the behavior of an individual at 2 A.M. at his desk in his own home may differ markedly from his behavior at 10 A.M. at his desk in his classroom. The statement about the observations, necessarily, must contain specification of condition.

What is to be observed is, however, crucial. No observation can be made unless man has arrived at some concept of it. For ages, man disregarded certain objects or their effects because he did not know of them or their behavior. He did not notice ultraviolet radiation, nor the fact that quartz reacted differently from glass to ultraviolet light. Nor did he notice electrical currents in the brain, or that some people can taste phenyltheo-carbonate. What is observed depends upon man's conceptual equipment to translate sensory experiences into the notion of a property or of a characteristic. The notion of property involves the commonness of observations as sensed by the observer.

The concept of a *property* or *characteristic* of an object or of a person is crucial in measurement. Every object has as many apparent, and different, characteristics or properties as there are different ways one can conceive of it. Physical objects have properties such as height, weight, color, shape, function, etc. Humans have not only the properties of physical objects, but also such other characteristics as intelligence, health, age, education, and personality. In science the attempt is made to observe a single property, characteristic, aspect, or trait of the object at a time.

The property observed, therefore, is dependent upon man's ability to conceive of it and, then, of his ability to observe it. Frequently, in making attempts to observe a conceived-of characteristic, the conception of the characteristic undergoes significant change. The adequacy of observation is a primary antecedent to the adequacy of measurement.

Statements about a property are empirical since they depend upon what is experienced. Such statements are called existential since they communicate what is observable by the senses or by extensions (machines) of the senses. In science, different observers must agree about the observa-

tion. This implies social acceptance about the fact of the observed phenomenon. If only one observer and no one else could make the observation, there could be no demonstration of its wider acceptability. In general, science demands a "reproducibility" of observations. Whenever conditions and methods are identical, the observations should be identical unless the object underwent some changes during the observation or subsequent to it.

A property, therefore, should be rigorously defined. It is usual to specify the conditions under which the observation can be made. In the observation of the tensile strength of steel, the specification of the operations of preparing the sample, of fastening it in the instrument, and of the instrument itself, are necessary to the definition of the property of "tensile strength." Similarly, in the observation of the aptitude of students for graduate study, the specification of the operations for giving the test, of motivating the student, and of the intelligence scale itself, may be necessary to the definition of the property.

Judges try to make estimates of some given property or characteristic of an object with little or no regard for any of its other properties or characteristics, and with little or no attention to its surrounding environment. The length of a table may be estimated as longer than some present, or remembered, or thought-of standard of length without any attention to other characteristics of the table; the judgment is made without reference to its color, width, height, wood, style, shape, or other property. Psychometrists try to make estimates of the "intelligence" of an adult with little or no consideration of his race, personality, or economic circumstance, and with little or no attention to his dress, the time of day, the furnishings in the room, and so on.

A property is some feature that is considered to be common to all objects of a given kind or class, but that need not be considered to be present in all objects whatsoever. From object to object in a given class, therefore, there will be some feature that is common to all the objects. It may be that the amount of this property will vary among the objects within the class, but regardless of other differences, there will be resemblance in the possession of amounts of the common property. Objects, of course, may have several properties in common. As already indicated, humans have properties of height, girth, and weight in common. Some judges may try to consider variations in these three characteristics independently; others may try to regard the combination of the three traits plus others under the *property name* of health. The property, thus, may be relatively simple and narrow, or it may be relatively complex and comprehensive. The property or trait, comprehensive or narrow, is attributed to the object by some person or persons.

The attribute, therefore, rests upon the direct or indirect perception of an observer. The ascription of a property to a class of objects, however, requires that differences in other characteristics be ignored, while the attention is given some main feature or quality. A property, broad or narrow, that can be seen, recognized, or classified only by a unique observer, however, cannot be considered a property capable of measurement or enumeration. In science, at least, there must be an agreement among expert observers that the property does exist, and such observers must agree in excess of chance upon whether a particular reaction does or does not demonstrate the property. (If different judges of the same incident cannot agree whether telepathy took place in the behavior they observed, no scientific analysis of the data is possible.) If judges perceive the same differences, then, the perceived differences may become the basis for classification and enumeration and, in some instances, measurement. The perception of resemblances and differences becomes the basis for measurement.

Empirical measurement depends upon the occurrence of a sense datum and its interpretation by an observer. Weight, even though measured by machine, nevertheless is ultimately related to the kinesthetic sensation of heavier and lighter. The machine—that is, scales—merely allows for a simple and relatively objective perception of the effects of weight. In this sense, the Geiger counter is a machine that enables the observer to perceive a specified class of effects by extending the range of human sensation. Other machines facilitate classification and comparison of objects by the control of systematic, chance, or erratic conditions, or by magnifying effects. These machines, therefore, allow more precise determinations of a particular class of effects. In the absence of instruments for extension of the senses, or in the absence of machines for the control of conditions, or in the absence of devices for the recording of more precise estimates, human observations are very liable to error. These sources of error include parallax, optical illusions, confusion of perceptions as in the size-weight illusion, biases, and sets. Instruments are a means for approximating more closely the property under observation. Instruments are steadily improved upon, so that the newer instruments make for closer and closer approximations to the measurement of a specific property. The object to be observed causes a variation in the instrument; the observer's sense impression leads to a recorded, objective, representation of the property.

But there is a limit to the closeness of approximation in the measurement of a property by an observer or by an instrument. In calorimetry it is well known that the temperature of the measuring device affects the temperature of the material under observation. In routine measurement, the observation is corrected by calculation for the influence of the instrument.

In this sense, the instrument affects the measurement. In intelligence testing the examiner can affect the score of the candidate. In individual testing, encouragement or its lack influences the child's performance. At present, however, there is no way of estimating a suitable correction for the rapport between examiner and child. Much of the attention of the physical sciences is devoted to the reduction of the interaction of the instrument upon the characteristic under observation. In the field of psychology, the growing emphasis upon social psychology illustrates the recognition of the influences of many factors upon perception.

Classification and Enumeration

Perhaps the simplest form of observation is to perceive that two objects are similar or dissimilar in their "response" to some situation. Eventually the observer notes that he is recognizing the same likeness or difference with respect to a great many object-pairs, and at this point he abstracts the similarity as a conceived property. He may then group objects which are similar in this respect into classes, using such dichotomies as short *vs.* not-short, heavy *vs.* not-heavy, or perhaps, intelligent *vs.* not-intelligent. The definition of the subclasses "intelligent" and "not-intelligent," or the subclasses "heavy" or "not-heavy," becomes the primary consideration for classification. The definition requires a specification of the basic distinguishing signs for judging inclusion within a class. Objects or persons then can be sorted into the class by a kind of "go," "no-go" criterion. Whenever observers can agree on the stigmata for class inclusion, then they can communicate to each other by using some class name to designate objects within the class. In this way, there may be agreement on the class "heavy objects," or "dull persons," or "basketball players."

All objects that "go" into the class are considered the same or equal under the convention that all other differences are disregarded; that is, all the objects that make a certain response are put together, even if other responses are not alike. Of course, within such a class as defined, further subclasses may be recognized on the basis of some characteristic common to the members of the subgroup but not to all the members of the class. Thus, the subclasses of "forwards" and "centers" may be found within the class "basketball players." If the definition is so clear and precise that observers can agree in placing certain persons into the class "basketball players," or certain basketball players into the subgroup "centers," these persons can be enumerated or counted. Class membership, therefore, rests upon an observer's acceptance of observable likenesses or resemblances so that each class member has the sign distinguishing it from all non-class members.

While enumeration assumes that the class is homogeneous, members of the class "basketball players" may differ among themselves in the quality of success as a basketball player. Certainly, coaches have to make judgments that some player or players are better (or worse) than others. The coach may define two mutually exclusive subclasses of good and poor players; he may even make three or four such classes. If, within the total class, the distinction of better or worse (stronger or weaker, heavier or lighter) is made, a comparison is intended. It is the comparison among objects that gives the first level of the relationship of subclasses to each other. The objective of scientific observation is to systematize a comparison so that existential statements about property and relationships within property can be structured in reference to the working conception of the property or trait or characteristic.

Ranking

Even after all objects have been placed into their respective subclasses, it may be possible to make a further existential statement indicating their relative order within the subclass. If coaches could agree upon the sorting, several consecutive subclasses of basketball players could be formed such as superior, average, and below average. As a matter of fact, always assuming the perceptual ability of the judge, there may be as many subclasses as there are players. In either case, the players would be ordered in terms of the property "success as a basketball player." In some instances objects may be ranked even though discrete subclasses cannot be distinguished. Although the hues of the solar spectrum, for example, can be ordered, they can be divided into the classes "reds," "yellows," etc., only by imposing arbitrary boundaries.

Observers can recognize differences in the relative amount of many properties. Shades of blue can be distinbuished in the range from the pure hue to a nearly washed-out white; apples can be graded on the basis of discriminable marks; and the scholastic successes of children are judged by their teachers. It is not unusual for a teacher to rank one student ahead of another in scholastic success. As a matter of fact, some teachers can arrange the students in a class in a rank order, so that Robert stands ahead of Jane who stands ahead of Elizabeth.

Ranking may be based on a judgment of some observed property, as when contestants are arranged according to beauty. Sometimes the ranking is based on direct comparison, as may be the case in a ladder tournament where one determines the order of basketball teams by permitting them to play against each other, pair by pair. A good example of ranking based on

direct comparison is the replacement series of metals. When zinc is dropped in a copper solution, copper plates out and the zinc dissolves. But copper will not replace zinc in a zinc solution. By successive experiments, it is found that a replacement order exists, beginning with gold, the easiest to replace: gold, mercury, copper, zinc, sodium. Such an order can be established without an understanding of the property. All that is necessary is to observe a hierarchical relation. For metals, a hierarchy exists, since sodium replaces all the metals listed before it, zinc replaces all those listed before it, and so on. Never is the order reversed, as it would be if copper could replace sodium. Sometimes comparison can be made but ranking cannot. If we observe a teen-ager order ice cream on many occasions, we may find that she consistently takes chocolate rather than banana, and that she takes banana rather than vanilla; but that when given the choice between chocolate and vanilla, she takes vanilla. Ranking is not possible unless objects possess the property under discussion (in this case, desirability) in increasing degree, so that one object has the property to a greater extent than any object below it in the series.

For instance, if problem-solving ability were considered to be a unidimensional property, individuals could be ordered according to the level of problem that each was able to solve. It would have to assume that the problems exist in a perfect order, and the individual could do all of the problems up to a certain one and none beyond it. Usually this requirement is not satisfied in educational and psychological measurement, which indicates that the property being investigated is a complex resultant of several simpler properties. Size is a property which does not permit perfect ordering. If size is defined by the operation, ability to pass through apertures of increasing area, some objects can pass through apertures of increasing area, some objects can pass through an aperture that apparently smaller objects cannot, because of differences in shape. We cannot arrange a miscellaneous collection of objects, and be sure that one object will pass through every aperture that an object nearer to the "large" end of the scale will pass through. Size, as a property, can be clarified by redefining it in terms of the unidimensional properties, such as length, breadth, and thickness.

The value of such observations of order is great. One of the earliest ordered observations involved the development of a method of communicating wind velocity. Lacking an instrument for measuring wind velocity, judgments were made about it. About thirty different subclasses were recognized from a calm up to a hurricane. Intermediate steps were gentle wind, breeze, and gale. For purposes of reporting, however, all wind

velocities that experts called "calm" were symbolized by the numeral 1; all wind velocities called "gentle wind," numeral 2, and so on, up to numeral 30 for "hurricane." The numerals are labels in an arbitrary hierarchical order. As a matter of fact, numbers 6, 7, 8, . . . 35 or 4, 8, 13, 15, . . . 47 could have been used to label the subclasses from "calm" to "hurricane." From these statements, one cannot say that two hours of calm would move a ship as far as one hour of gentle wind.

Two objects may be one numeral apart, and, yet, that difference may be either very small or very large. Suppose a teacher were to rank the heights of the members of a class of thirty-three youngsters. It is reasonable to expect substantial reliability in such ranks. Pupils in ranks fifteen, sixteen, and seventeen, as measured on a stadiometer, would probably differ little in height, but pupils in ranks one and two and those in ranks thirty-two and thirty-three would differ much more in their measured heights. The numbers assigned to the rank orders give useful information about relative position, or direction of differences, but they do not allow inferences of absolute amount or of the difference or ratio between the amount of one object and the amount of another.

Ranks are not fixed. If thirty-two pupils were ranked in height, the set of numerals first, second, third, . . . could be put into one-to-one correspondence with Robert, Jane, Elizabeth, and so on, to John who is paired off, let us say, as thirty-second. If some independent observer arranged the thirty-two students in the same order, there would be a feeling of confidence about the number assigned to each pupil. There may be students, in existence, however, who possess amounts of the quality "height" that would require some of them to be placed ahead of Robert, or between Jane and Elizabeth. These new students would change the rank order and the sequence of numbers assigned. The ordinal numbers merely mean that 1 (for first) has a rank higher than 2 (for second) and that since 2 has a rank higher than 3, 1 must rank higher than 3.

Ordinal arrangement can be developed into a more complete observation if it is possible to assess how near to each other the objects are in the property in question. If several objects can be said to be equally spaced, in terms of the degree to which they possess the property in question, the distance between any two of them may be taken as a linear unit of measurement. To describe the interval between two notes in a melody, the tone or step is a suitable unit. The difference between C and D is accepted to be exactly that between D and E. Any interval can be described in terms of the number of submultiples of the unit from one end to the other of it; for example, C to G is three and a half steps or seven semitones. In the

measurement, a distance between two sounds (an octave, a tone, a cycle) is chosen as a unit. The particular unit (though not the nature of the unit) is necessarily arbitrary, and is chosen for reasons of historical accident, convenience, or invariance (as in the use of the cadmium line in the spectrum as a measure of length). If one has a measuring device on which successive equal intervals are in some way "marked" or recorded, one may compare an object with the standard. The observer reads the number of units needed to "equal" the test object. The comparing and reading are visual in the case of length, but a device such as a pitch pipe can be used as a standard for tone, and the comparison is made by the appropriate operation of listening and comparing.

The particular unit chosen must have relation to the perceived property. In measuring income, the dollar is a suitable unit if the property under study is the money each person receives. In those terms, appropriate in an accounting office, the gap between \$1,000 and \$2,000 is the same as the gap between \$5,000 and \$6,000. But if the income is conceived in terms of goodness of living permitted, the count of dollars does not linearly represent the increase in reward. The increase from \$1,000 to \$2,000 may represent the change from penury and misery to the bare ability to obtain necessities; the increase from \$5,000 to \$6,000 may represent the change from a comfortable level to the beginnings of luxury, or it may represent the comfortable level with an opportunity to safeguard it with savings. No satisfactory unit for measuring income, conceived as reward, has been developed.

In some fields, an increased understanding of the relation of the perceived property to other important properties has led to increasingly satisfactory units. In estimating hearing, a primitive measure was the distance at which a child could hear a sound. In the watch test, a ticking watch was brought closer and closer to the child, until he could just hear it. This is a simple experimental device for arranging children in order, in terms of keenness of hearing. The score on the test (distance) could be taken as a scaled measurement only if observers agree that children who hear at 20 feet, 15 feet, and 10 feet, respectively, differ from each other by equal degrees of keenness. Another method that might be used is to produce varying sounds at standard distance. A primitive scientist might have dropped increasing weights on a drumhead, and noted the smallest weight that would produce a sound the child could hear. In modern times, these crude methods can be replaced with an electronic technique. The intensity of the stimulus might be measured by the pressure of the sound waves emitted by the resonator, the unit of pressure being the dyne, which is a

unit abstracted from the falling of weights. Then hearing might be scored in terms of the number of dynes per square centimeter. Such a scale is not linear. If a sound of 20 dynes per square centimeter force is compared with another of 40 dynes, an observer asked to locate the sound halfway between the two in loudness will set the resonator not at 30 dynes, but at 28, the geometric mean. As the experiments of Weber and others show, the equal intervals of the hearing scale have a logarithmic, not a linear, relation to units of force. Similarly, it can be shown that since the force of the sound decreases proportionately as the *square* of the distance from the source, the distance measured in the watch test is not a linear measure of sound. From many studies of the nature of hearing, of what it means to say that one sound is twice as loud as another, has come a unit of measure known as the decibel. The decibel is a logarithmic function of the power of the sound; the sound that is 1 decibel louder than another is ten times as strong in terms of dynes, and a 2-decibel difference corresponds to a ratio of 100 to 1 in sound energy. The decibel is a unit on a scale having equal-*appearing* intervals, linear with relation to loudness as judged by observers. The establishment of equal-appearing intervals rests on the agreement of judges in setting the tone halfway between two standard tones.

The linearity of a scale, it should be noted, is linearity with relation to some judgment or some operation. The householder can use number of pounds as a unit in buying coal, because the weight is roughly in terms of B.T.U. per dollar. Although the ton is a useful unit, the buyer can judge value more accurately with a unit such as the B.T.U., which indicates the amount of *heat* he will purchase, rather than with a unit such as *tons* which indicates the amount of *coal* purchased.

Because measurement always deals with a property which has been previously perceived, and is to some degree familiar, there is a danger of believing that the "obvious" unit is in some way peculiarly right for measuring the property. There is a compatibility between the notions of the height of persons and the length of a measuring stick that makes the inch an appealing unit for measuring growth. In fact, common sense rejects as absurd any suggestion that the footrule does not provide a scale of equal intervals for height. Yet height can be measured in years rather than inches! All that is required is the median height of six-year-olds, seven-year-olds, etc. Then, if we have a record of these data, we may compare a given child with the standards and report that he has seven-year-old height; or, perhaps even more precisely, that his "height age" is 7.2 years. The only justification for this seemingly roundabout and "un-

natural" measuring scale is that in some studies one year of growth makes a unit for a scale of equal intervals, whereas the inch does not. An analogy is found in the measurement of the maturity of trees. Once, the largest, or thickest, tree of a species was thought to be oldest. Now, we rely on a count of the annular rings as a measure of age, even though the rings are unequally spaced on a footrule.

Very frequently, scientific observation is concerned with making statements about differences in amount. It is true that the observer using the scale for wind velocities can say that there is an observable difference between a calm and a gentle wind, or between a gale and a hurricane. The scientist, however, would like to state how much difference there is between adjacent or different velocities. To make such statements of difference, however, implies that juxtaposition, for example, putting two weights in the same pan of a balance, will yield a weight equivalent to their sum, or that putting two objects on top of one another will yield a height equal to their sum. Height is such a property. Neither a 36-inch child nor a 30-inch child can reach the jam jar on the top shelf; but their heights may be added, when one stands on the other's shoulders, so that they do reach the shelf. Psychological properties rarely permit addition. If Sarah knows 200 French words, and Jane knows 250, their total vocabulary may be only 320 words, since many words appear in both vocabularies. Two readers, each capable of reading at the third-grade level, cannot by combining read a book of sixth-grade difficulty. Very few of the properties with which education and psychology are concerned can be measured by additive scales. The physicist, on the other hand, is able to measure with the additive scales for length, time, mass, volume, electrical resistance, heat, and several other properties. Properties such as density and temperature, which are not themselves represented in additive scales, may be derived from, and measured by, combinations of properties for which there are additive scales.

In an operationally defined additive scale, one may make meaningful statements about the *amount* of a difference. The difference from 48° to 32° is accepted as equal to that from 16° to 32°, because the units of the temperature scale are equal by definition. But the difference cannot be interpreted as 16 times a unit of temperature, since there is no invariant unit of temperature permitting addition. On the other hand, chronology is a scale for which the additive principle holds. It is meaningful to say that two events occurred ten years apart, since the year is an additive unit, demarked by the earth's rotation about the sun. "One year later" is an additive statement, marking off the point in time one revolution beyond the

base point, just as "one foot above his head" implies the addition of lengths. In the best-developed additive scales, it is operationally demonstrated that *units* of the property may be added.

Even additive scales fall short of complete measurement until an absolute zero is established so that the investigator can answer the question "How much?" Additive scales permit measurement of differences, but cannot report absolute magnitudes unless they can measure the difference between the object and a zero point. For example, the chronological scale permits comparisons of times, but the number "1902 A.D." does not state how much time elapsed prior to the designated event. The birth of Christ is an arbitrary zero, useful as a bench mark but meaningless as a number, since time did not begin at that instant. Because these numbers are not based on zero, multiplication and division of them is not meaningful. 2000 A.D. does not represent twice as late a point in time as 1000 A.D.; nor did the age of the earth double from year 1 to year 2. In order to make statements of absolute magnitude, a ratio scale is required.

The zero for a measure of a property is the amount that is conceived to represent "just not any of" the property. Zero illumination is total darkness. Absolute zero for intensity of sound is the point where movement of the air (or other conductive medium) ceases. Heat, on the other hand, had no absolute zero until relatively recently. Even when the temperature is 32° F. or 0° F., objects have not reached the point of no-heat-at-all. A ratio scale can be established when an additive property has a functional relation to a phenomenon which does have an absolute zero. When heat was related to thermal motion, and zero heat was equated to the cessation of thermal motion, the ratio scale for heat became a possibility. One object could justifiably be said to have twice as much heat energy as another. A scale may have an absolute zero and yet not permit measurement, if the additive property has not been proved. Absolute zero error in rifle marksmanship may be readily conceptualized in terms of zero deviation from the bull's-eye, but marksmanship itself has no scale of additive units for measuring accuracy.

The notion of absolute measurement was considered by Gauss around 1830. The acceptance of the concept has been slow. In physics, studies of the conservation and dissipation of energy lead to Thomson's absolute scale of temperature about 1848. The extension to electrical phenomena occurred around the same time. In physics, a hypothetically observable but nonexistent phenomenon (just not any distance between two points, just not any deflection of an inertia-free galvanometer) is related to the zero point of

the interval scale. In some scales now used without a true zero, a zero may yet be defined. Conceivably, for example, measures of radioactivity might set the age of the earth closely enough that absolute times could be used for the earthly chronology (not for eternity). In the social and psychological fields, unfortunately, outside referents have not been conceptualized. It will perhaps not be possible to establish true zeros for behavioral properties such as prejudice or sociability. It is unlikely that a measure will be developed which permits the statement that child A is twice as good as child B in ability to solve problems. Nor can a handwriting sample of zero legibility be defined, since no matter how poor the sample, there is always the probability of a worse one.

Thus far, the emphasis of this chapter has been upon observations, and upon the human ability to classify observations. In essence, classification depends upon the human recognition and conceptualization of some uniformity among observations. The greater the exactness of the formulation of the concept of the characteristic of the property, the greater the abstractness of it. In this sense, the concept of a property involves an abstraction from the concrete, or perceived, observations.

The concept of a characteristic may be such that things, people, or events may be considered to have it or to lack it. At this level, objects, for instance, may be separated into two groups, the have's and the have-not's. The concept, however, may be one that recognizes differences in degree or in amount. The concept of a property that recognizes differences in degree has been called an *intensive magnitude*. An intensive magnitude represents the kind of property whose observations can be arranged in a recognizable order, but where two observations or events cannot be added. For instance, an observation of eighteen arithmetic examples for John and of twelve for Jane is not thirty for both. Another illustration of a property of intensive magnitude is E. L. Thorndike's achievement scale in drawing (6). The plates of the drawings can be arranged in a recognizable order from best to worst performance, but the value of any two drawings cannot be considered equal to any third.

The property of weight, however, is such that objects can be arranged in a *definite* order, that there is a limit which means "just not any" weight, that the sum of any two would equal a third. Such a property has *extensive* magnitude. The distinction between intensive and extensive magnitude has been given in the table on the following page.

An intensive magnitude refers to observations which can be placed in order and for which the order is transitive and asymmetrical. First, it must

Intensive	*Extensive*
1. Recognizable order	1. Definite order
2. Indeterminate zero, i.e., zero is arbitrary or normative	2. Absolute zero, i.e., zero is "just not any of"
3. Has some relation (not perfect) to an external criterion	3. Has an invariant relation to an external criterion
4. The sum of two substances of different values *is not equal* to that of a substance with the value of the numerical sum of these values	4. The sum of two substances of different values *is equal* to a substance of the value equal to the numerical sum of these values

The concept of property distinguishes between extensive and intensive magnitudes.

be clear that a class of objects B $(i = l \text{ to } n)$ have a common property. (I) If the objects have a common property, then either

$$B_i = B_j, \text{ or}$$
$$B_i < B_j, \text{ or}$$
$$B_j > B_i.$$

This is the statement of asymmetry. And (II) if $B_i < B_j$, and $B_j < B_k$, then $B_i < B_k$. This is the statement of transitivity.

It must be noted that the equal sign of (I) and the less sign of (II) refer to *observed* relationships. The statement of equality, moreover, is not necessary to the placement of all objects in an asymmetrical and transitive order. Campbell (*1*) has shown that in genealogy, the construct "=" does not exist. In such a situation "<" is read "is descended from." Yet the line of descent can be ordered in time.

The Mohr scale of hardness merely indicates that, in terms of the expressed relation of scratching as evidence of hardness, minerals can be placed in "scratch order." There is no evidence that the hardness of a diamond is any multiple of the hardness of any other minerals.

In order to express the concept "multiple," the property under observation must be capable of physical addition. Properties that can be physically added are called "extensive." It is necessary to demonstrate that in addition to the conditions of asymmetry and transitivity, the following conditions also apply:

III. if $B_i = B'_i$ and if $B_j = B'_j$, then $B_i + B_j = B'_i + B'_j$;
IV. if $B_i + B_j = B_k$, then $B_j + B_i = B_k$;
V. if $B_i = B'_i$, $B_j = B'_j$, and $B_l = B'_l$, then $(B_i + B_j) + B_l = B'_i + (B'_j + B'_l)$;
VI. if $B_i < B_j$, then $B_i + B_k < B_j + B_k$ if $B_k \neq 0$.

It must be understood that the property B is common to all objects B_i

and the operations of addition and of combination correspond to the physical process as does the joining of two sticks together or the putting of two weights in the same pan.

Since measurement involves the assignment of numerals to properties, it is necessary to understand the formal conventions about numbers. The formal conventions among numbers are essentially arbitrary. Each number system represents a certain set of formal and logical relations among the symbols. Numerals may be nominal, ordinal, and cardinal. Nominal numbers are mere symbols for identification without further property. Ordinal numbers refer to the order of objects as first, second, and third. Cardinal numbers refer to enumeration of objects which can be counted regardless of order.

Just as words have a variety of meanings, so do the numerals. Much of the difficulty would be reduced if there were different symbols representing a nominal number, a rank order, an interval scale, or a ratio. There is little likelihood that four such sets of symbols will be developed. The scientist must be ever on his guard to recognize from context whether the 1, 2, . . . refer to nominal, order, count, or ratio.

The numeral assumes that all objects or events given the same designation, whether a name or a number, are equivalent with reference to the named property. Nominal scales merely name, by number, letter, or other symbols, all objects having the same property. Rank-order or ordinal scales assume that the objects or events have a common property and that the objects differ in the amount of that property. Essentially an ordinal scale is concerned with the arrangement of events into a hierarchy from least to most. Rank order scales have been used in the estimation of the velocity of winds, the hardness of minerals, the conduct of pupils, the taste of foods, and so forth. Interval scales employ some invariant referent, such as a degree in temperature (defined as 1/100 the difference between freezing and boiling of a specified substance), or one j.n.d. (just noticeable difference) in brightness. For an additive scale, the units must be operationally addible. In ratio scales, the lower end of the scale is anchored at the point corresponding to "just none" of the property being measured.

A numeral may be used as the name of a particular object, as the name of a particular set or subclass of objects, as the enumeration of all of the objects in a set, as a rank of an object in an ordered sequence, as the amount of some property. The number 428, for instance, can refer to a particular safe deposit box, or to all eight-inch files of model 42, or to the fact that the shipping clerk had on hand just 428 hammers; it can also refer to a student's scholastic position in his graduating class, to the score a

student reaches on an intelligence test, to the melting point of an alloy of zinc, to the length of a wall in inches, and to the weight of a hog in pounds.

When a numeral is used to tag a safe deposit vault, or to give a code number to a salesman, it serves no other purpose than that of identifying a specific object. When, however, it is used to designate all model 42 eight-inch files as 428, it recognizes the equivalence of each model 42 eight-inch file with every other. The number 428 is assigned to the specific deposit box or to the salesman or the files in an arbitrary and capricious manner. Some record of the intended identification must be available for recognizing the reference.[1] The key is necessary to make the numbers intelligible in communication.

The fact that there are 428 objects classified as hammers, however, is a different kind of numeral. It is that number of the series of natural numbers, beginning with 1, that is paired off with the last member of the class of hammers. The convention adopted is that the sequence of natural numbers 1, 2, 3 . . . exists and that all objects of the class may be considered to have the property of hammerness or to lack that property. The class or set is defined by any property which each object considered must either have or lack. Enumeration is a specific operation by which the sequence of natural numbers is paired off with each unit of the class "hammers" until the supply of hammers is exhausted. The number that is assigned to the last hammer constitutes the count.

If the first natural number is paired with the first object, then, the natural number paired with the last object is the cardinal number representing the count. If the symbols refer to ordinal sequence, the numbers can have meaning only with reference to the set of ranks used. Any addition of new members to the class changes the interpretation of the ordered number. Pure rank as eighth means that if the pairing begins with first, then between first and eighth are six other ranks. The interpretation of eighth-in-a-sequence-of-eight and eighth-in-a-sequence-of-five-hundred obviously cannot be the same.

The numerals assigned to the temperature scale or the length scale, however, do not refer to discrete properties. They refer to continuous property. The numeral 1 can no longer be considered fixed and unvarying. The numeral 1 refers to an approximation such as $1 \pm a$ is the variation in reading the results of a juxtaposition of scale and object. As a matter of

[1] In some instances, the numeral may be somewhat more meaningful. For example, Washington, D.C., may be assigned the number 226 because it is uniquely identified as that many miles south of New York City on the Pennsylvania Railroad.

fact, unless the zero is established, the amounts can refer only to additions of units. When the zero of the scale has been established, then ratios can be reported such as $\frac{A}{B} = K$.

Stevens (*5*) has made a succinct classification of scales of measurement, see Figure 60. His table shows the relationship between the formal concept of numbers (that is, scales) and the existential concept of property (that is, basic empirical operation) together with the mathematical group structure of the numbers, their permissible statistics, and typical examples. This table summarizes the relationship between number or symbol and property of observation.

Measurement as Successive Approximation

The concept of a scale rests fundamentally on the ability to recognize some property or attribute, and to order objects with relation to sensed differences in the property. The first bases for recognition of the property may be imperfect, due to the limitations of the observer or to misconceptions about the property itself. Judgments of the pitches of tones can be made which will permit ordering the tones into a series. But when careful experimentation shows that pitch, as a sensed quality, varies with both intensity and frequency of vibration, a more satisfactory measurement of pitch becomes possible by expressing it in these dimensions, which are objectively measurable. A relatively crude measure of physical development was over-all stature. While stature did increase as development progressed, many children who were relatively immature were also relatively large, so that this measuring scale was only an approximation. A later and more adequate indicator of skeletal maturity was the ossification ratio, an index of the extent to which calcium had been deposited in certain parts of the wrist. This became a more satisfactory basis for estimating maturity, although this is far from ideal. Moreover, it yields only an ordinal scale of development, and further refinement of the concept is required before an interval scale or a ratio scale will be possible.

In developing a measurement, many changes are made in the conception of the property being measured. The first estimate may be crude and impure; thus Binet used teachers' judgments of intelligence as a starting point even though he knew these judgments contained error. By determining the relation of the property measured by the scale under study, to other properties or other measures, the investigator clarifies his scale. Weight no doubt was first conceived through kinesthetic sensation, and scales of measurement were developed. When observations showed that weight changed with position on the earth's surface, and when Newtonian theory provided an

Scale	Basic Empirical Operations	Mathematical Group-structure	Permissible Statistics (invariantive)	Typical Examples
NOMINAL	Determination of equality	Permutation group $x' = f(x)$ [$f(x)$ means any one-to-one substitution]	Number of cases Mode Contingency correlation	"Numbering" of football players Assignment of type or model numbers to classes
ORDINAL	Determination of greater or less	Isotonic group $x' = f(x)$ [$f(x)$ means any monotonic increasing function]	Median Percentiles	Hardness of minerals Quality of leather, lumber, wool, etc. Pleasantness of odors
INTERVAL	Determination of equality of intervals or differences	General linear group $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation	Temperature (Fahrenheit and centigrade) Calendar dates "Standard scores" on achievement tests (?)
RATIO	Determination of equality of ratios	Similarity group $x' = ax$	Coefficient of variation Logarithmic transformations	Length, weight, force, etc. Pitch scale (mels) Loudness scale (sones)

Measurement is the assignment of numerals to objects or events according to rule. The rules and the resulting kinds of scales are tabulated above. The basic operations needed to create a given scale are all those listed in the second column, down to and including the operation listed opposite the scale. The third column gives the mathematical transformations which leave the scale form invariant. Any numeral, x, on a scale can be replaced by another numeral x' where x' is the function of x listed in column 3. The fourth column lists, cumulatively downward, some of the statistics that show invariance under the transformations of column 3. (For further details, see S. S. Stevens [5].)

FIG. 60.—A CLASSIFICATION OF SCALES OF MEASUREMENT

(Reprinted by permission of S. S. Stevens, Harvard University)

explanation of the property in terms of attraction, the concept of mass became available as an alternative. Whether mass or weight is the more meaningful property to measure depends on the purpose of the experimenter. But determining the relationships and distinctions between these two similar properties adds to the meaningfulness of each of them.

Perfection in the measurement of *intelligence* or *achievement* is not likely. Physical measurement, too, fails to realize perfection. Despite improvements in instrumentation, despite control of conditions, despite corrections for intervening phenomena, physical measurement still has a residual uncertainty. The classical belief that improvement and corrections will converge to an ultimate perfection is fallacious.

There will always be an indeterminate interaction between the observed and the observer. Psychology is not alone in its recognition of the interdependence of the property and its measure. A close analogy to the Heisenberg principle of uncertainty, or better, of indeterminacy, applies to the measurement of the speed of an electron and the appraisal of the subject-matter mastery of a student. The process of measuring the student's knowledge itself changes the level of his ability, by stimulating study, inducing emotional strain, etc. There is a limit to the precision of the phenomena measured. Beyond that limit, improvement in the measuring device will not lead to reduction in the amount of indeterminacy.

The recognition that all measurement ultimately has a residual of uncertainty affects the certainty of causality. When will just no thermal movement take place? When will the behavior of a child be free of the influence of the recorder? In place of the more conventional statement of cause and effect, modern concepts of measurement substitute statements about the probability of an event, or of a property. The statistical notion gives a probable value for the relation of an organism to its environment, or for the interaction among observers. The statistical concept regards all elements as equivalent. The idea of equivalence of elements neglects the specific nature of the elements that make up the property. Statistically, the property is measured in terms of aggregates of elements, of averages, and of variations. The measurement of the smallest quantity of light that can be seen is a statistical determination, not an absolute determination.

The refinement of a measurement may be illustrated by the problem of evaluating a person's knowledge of words. Word knowledge is a familiar notion. A person's word knowledge may be inferred from his speech or from his writings. An apparently superior basis for judgment is to ask the subject to define a series of words. Then, the count of the number of words he is able to define may be considered as word knowledge. But there is a serious problem of equivalence to be faced. If the list

of words is poorly chosen, the scores on the list may not correspond to other judgments about word knowledge of the subjects, and so be unacceptable. A common criterion, representing what is normally meant by word knowledge, is the ranking assigned by a teacher who judges the verbal competence of the children. If the number of acceptable meanings given by the children in response to the list of test words corresponds with the rank order given by the teacher, the conclusion is that the number of word meanings may be considered an ordinal scale for measuring vocabulary knowledge.

One requisite for the measurement is the interpretation of responses considered as evidence of acceptable word knowledge. When for "dress," the child points to his eye, or makes a pushing motion, or says "you eat it" or "it's a kind of automobile," the child would not be credited with knowledge of the word; when, however, he gives any acceptable indication that he has a meaning of dress, he is given credit by counting it as "one word known." Very young children tell what a word means by telling what the thing is used for, or by pointing to the thing, or by demonstrating. More mature children, however, give meanings in terms of synonyms, or as a genus-species relationship, or in terms of explanations. For instance, the direction may be, "What does *dress* mean?" The responses may range from pointing to a dress, demonstrating how to put on a dress, to saying "You wear it," "Clothes," "A woman's clothing," "To ornament," "A nurse or a doctor caring for, or bandaging, a wound."

The definition of dress in terms of genus-species or a synonym is usually regarded as a somewhat superior performance to that of pointing or demonstrating. Yet, as a first approximation to the relative vocabulary knowledge of a child, the convention is that any indication of meaning is "equally" acceptable. By this convention, pointing to a dress or saying "caring for a wound" are given equal credit, despite the fact that there is an apparent difference in the content and process of the response. The content for the young child will usually be of "dress" as a garment, whereas for the adult it may go beyond to more specialized and recondite meanings. Furthermore, the process of definition will range from demonstration to statements of use for young children; but, for more mature adults, the process may be of explanation, seeing equivalences, deducing relationships, and the like.

The child, basically, is given a credit of 0 if he does not show any grasp of any meaning of "dress"; and a credit of 1 for any indication of its meaning.

The units of word knowledge are not equal in another sense. Knowledge of the meaning of "dress" is not equal to knowledge of "entropy" or of "collyrium." While there may be a child who does know "entropy" without knowing "dress," common sense tells us that this happens rarely. Not

only are the units not equal in the sense of word rarity or reconditeness, but also the units are not equal either in the level of the different meanings known, or the process by which they were acquired and expressed.

The device of crediting definitions for words as zero and one for each subject gives a score which is an approximation to his general word knowledge. The number of words credited neither indicates which words are known, nor what meanings of those words are comprehended. Moreover, the credit does not give any evidence of the kind or degree of mastery the child has. As a consequence, the score in terms of number of words adequately defined is a mixture of vocabulary range, vocabulary precision, and conceptual maturity.

The crude count of words known, dependent as it is upon an arbitrary standard of acceptance and upon a finite number of words, is, nevertheless, an estimate superior to the more usual rating, or judgment, of vocabulary competence. At least within limits, the number of words known, among a specified sample of symbols, yields an objective count. Such objectivity tends to make the observer neutral in the process of measurement. A count of words, arrived at by some standardized set of operations, will, in general, be freed from the contaminations of the observer's preconceptions and biases.

Crediting a word as known, however, is a statistical operation. The response the individual makes at a given moment does not necessarily represent all his knowledge nor the minimum of his understanding. It is merely what he happens to state at the time. Moving from individual to individual emphasizes the statistical basis of the appraisal of word knowledge. While each individual may get credit for the word, the process may differ from person to person. Recently, Feifel (*2*) showed that although psychotic and normal individuals obtained credit for each of the first ten words of the Terman list, the normals more often chose the synonym types of response whereas the psychotic tended to give definitions more frequently in terms of use, demonstration, and illustration.

The property of "vocabulary knowledge," or, more literally, "words defined in an acceptable fashion," is not invariant from person to person nor within an individual from time to time and from word to word. The units for reporting the property lack objective equality (or even apparent equality). The process lying behind equal amounts of credit will usually differ from word to word and person to person. Furthermore, the unit itself cannot be isolated as can the meter, in the form of a meter-stick; and there is no zero signifying just no word knowledge. Nevertheless, despite the imperfect units for counting word knowledge, such a count does give genuine information about the relative status of individuals.

This interest in a count which is recognized to be crude arises because the count is associated with other properties of importance, such as mental development with increasing age, school progress, and ability to read materials of varying difficulty. The "score" on a vocabulary test, even though it is an expression of rank on a scale which is not unidimensional, even though it is on a scale of unequal intervals, and even though it is not on an additive scale permitting sound arithmetical comparisons, is still a useful approximation. It does permit practical judgments, with greater accuracy than would be possible without a measuring procedure. When the basis of the scale is questioned, and it is shown that different processes lead to equal credit, the investigator may return to the original perceived property and attempt to purify it. Thus, an attempt to distinguish between "synonym" responses and "use" responses is a first step toward developing a new measuring procedure, which may eventually be more precise and more useful.

Many test makers and test users believe somewhat naïvely that if the number of words is transformed to a standard score, that is, whenever a raw score is referred to some average in terms of the measure of group variability, that true measurement is achieved. Linear transformations of raw scores $z = \frac{X - M}{s}$ will not make the raw score something it never was. It is true that the linear transformations make for an easier interpretation of the probability of attaining a given score (or standard score); or, they allow the comparison of relative position on one test with relative position on any other tests similarly transformed. Equal standards scores, however, do not make for equal units of a measured property.

If the measurement is not based upon an unvarying unit with scale form of equal intervals, mathematical manipulation will not provide the unit. Reasonable operations, such as absolute scaling, may clarify the primitive unit, yet they cannot produce a unit with a true zero or a scale of equal intervals.

Explanation as the End of Measurement

The conventions of test construction all too frequently confuse the understanding of social and psychological properties and traits by attempting to make the field conform to the standards of mathematics. Insofar as mathematical models correspond to the realities of a given field, there is a justification for their use. The primary concern of measurement, however, should be *for* an understanding of the entire field of knowledge rather

than *with* statistical or mathematical manipulations upon observations.

Knowledge will be advanced by recognizing what the empirical methods of measurement ignore. In the early development of a field of knowledge, main influences are evaluated. It is only after the main actions or effects are understood that subsidiary and secondary effects may be appraised. In the illustration about word knowledge, the main effect of vocabulary as related to age is appraised empirically. The secondary effect of response form in relation to age or in relation to personality structure is just beginning to be understood. The aim of measurement must ever be the explanation of, or the meaning for, observed phenomena.

In achievement testing, especially, the credit for the number right is related to the quality of the teaching process, the method of studying and learning, the intelligence of the subject, etc. In the days of the Army Specialized Training Program, reasonably homogeneous groups of individuals were required to take apparently similar courses. The results of the achievement tests, however, differed significantly from school to school. The inference may be that different modes of instruction (among other factors) made for differences in achievement. It is important to recognize what is hidden in the credit for a given achievement exercise. Different processes (for example, intelligent generalization or memory) or different methods (for example, drill or understanding) may each get the same credit. Even though the achievement exercises initially are selected to give widely spaced steps in difficulty, the step positions will not remain invariant under different methods of teaching and learning. The primitive unit of the achievement test is still an acceptance of descriptive equality, in which each problem regardless of its difficulty (and any other factor related to difficulty) is given an all-or-none credit.

The methods of identifying properties by statistical operations may do much to clarify the constructs. Yet, the general methods of factor analysis yield, at best, very rough approximations. By now, the psychologist deals with such components as space, word, and number and such components as reasoning, perception, and memory. As restatements of the data, such components may do much to clarify the nature of the data the psychologist uses. But every bit of reasoning must apply to something, whether it be an object or a symbol classifiable as space, number, or word. Or, reversing the argument, every aspect of space must be acted upon by an individual either in reasoning or in perception, or in memory. The identification of the factors as properties neglects the interaction between process and content. At best, such an interaction is approximated statistically; it is not identified with great precision.

The nature of measurement in the psychological domain is such as to

give a kind of primitive unit which can be counted. The counts, however, are arbitrary, depending upon the sample of items to which an individual is exposed and upon the sample of processes and reactions such items elicit in the individual at a given moment. Such counts can be put in recognizable order, and it may be assumed with limited confidence that individuals are actually arranged in terms of the property in question, in the same order as the counts.

In measurement, therefore, the line of development will be from the concept to its scientific estimation, and then to an explanation of its relations to other properties. The nature of the units, of course, does constrain the amount and kind of possible generalization. In the early development of a field, the clarification of the concept is dominant. The refinement comes by successive approximations. Absolute measurement, such as the physicist has developed, is an unlikely hope. The physicist recognizes the indeterminacy of his absolutes; the psychologist, too, must be aware of the uncertainty of his units and processes.

The psychologist's measurements are primarily empirical in the sense that they depend upon the social acceptance of observed differentiations. The observed differentiations, however, are dependent upon the social and cultural setting in which they are observed. It is a commonplace to remark that the objective of "habituation" of skills and knowledges at the turn of the century has given way to an entirely different set of goals.

From another point of view, moreover, the operations themselves change in meaning. As an example, let us consider the data collected by Jones and Conrad (*3*) with reference to the Army Alpha Examination. They administered the Army Alpha to practically all persons from ten to sixty years of age residing in seven different New England villages. On the basis of the result, they demonstrated a regular increase in total score from age ten to about age twenty. After age twenty through to beyond sixty, there was an average decrement in score. If "whatever the test measures" be called "intelligence," then "intelligence" is not the same thing for children and for adults. The average adult gets about half of his total score credits from two subtests of the Alpha: Opposites (vocabulary) and Information. The children, on the other hand, get but 10 percent of the total score from these two tests. This observation, of course, could not have been made in the absence of an instrument like Alpha. The value to science in establishing that intelligence test items are not homogeneous over time is very great. The response to each item may involve a different response and a different motivation. The process by which a person solves a novel task for the first time is not the same process by which he responds to it the second time.

Each of the separate subtests of Army Alpha has its own characteristic curve of increase, then stability, and finally decline. This suggests that the property "measured" is lacking in homogeneity. The criticism of the Stanford revision of the Binet-Simon scale is essentially that the elements in the scale at the earlier ages are not like those at the adult ages. Recently, R. L. Thorndike (7) traced back the intelligence test of candidates who had taken the Scholastic Aptitude Test for college admission. He found the records of individuals, for instance, who had taken the Stanford-Binet from one to twelve years previously. The correlations between the S.A.T. and the Binet taken from twelve to eight years before were around .35, whereas those taken from one to five years before were about .70. The earned credits on the examination are not equally predictive of a single criterion at or near age eighteen. Either the argument must be that the test is not a homogeneous array of items, or that the property is not invariant in time, or both.

Kenneth Norris (*4*) has followed up the achievement of adults. He administered the Stanford Achievement Test to adults out of school anywhere from one to thirty years. Here the performances, again, showed great variability. In general, arithmetical reasoning declined with remoteness from school. The verbal skills remained, on the average, unchanged.

The ultimate criteria of "adult status" or of "life use" are themselves nonhomogeneous composites. The prediction of a "nonhomogeneous" composite from a single or a set of "nonhomogeneous" measurements will, at best, yield approximations. The items of the test or the criteria have a residual ambiguity no matter how carefully they are "item-analyzed." Hence, the more immediate the criterion, the greater the likelihood of immediate homogeneity. Within relatively narrow time spans, the property may be considered, in an actuarial sense, homogeneous.

Selected References

1. CAMPBELL, N. R. *An Account of the Principles of Measurement and Calculation.* London: Longmans, 1928. 295 pp.
2. FEIFEL, HERMAN. "Qualitative Differences in the Vocabulary Responses of Normals and Abnormals," *Genetic Psychology Monographs,* **39**: 151–204, 1949.
3. JONES, H. E., and CONRAD, H. S. "The Growth and Decline of Intelligence: A Study of a Homogeneous Group between the Ages of Ten and Sixty," *Genetic Psychology Monographs,* **13**: 223–98, 1933.
4. NORRIS, K. E. *The Three R's and the Adult Worker.* Montreal: McGill University, 1940. 213 pp.
5. STEVENS, S. S. "On the Theory of Scales of Measurement," *Science,* **103**: 677–80, 1946.
6. THORNDIKE, EDWARD L. "The Measurement of Achievement in Drawing," *Teachers College Record,* **14**: 345–82, 1913.
7. THORNDIKE, ROBERT L. "Growth of Intelligence during Adolescence," *Journal of Genetic Psychology,* **72**: 11–15, 1948.

15. Reliability

By Robert L. Thorndike
Teachers College, Columbia University

Collaborators: Lee J. Cronbach, *University of Illinois;* Edward E. Cureton, *University of Tennessee;* Truman L. Kelley, *Harvard University;* Albert K. Kurtz, *Pennsylvania State College;* Marion W. Richardson, *Richardson, Bellows, Henry & Company;* L. L. Thurstone, *University of Chicago*

Whenever we measure anything, whether in the physical, the biological, or the social sciences, that measurement contains a certain amount of chance error. The amount of chance error may be large or it may be small, but it is universally present to some extent. Two sets of measurements of the same features of the same individuals will never *exactly* duplicate each other. In some cases the discrepancies between two sets of measurements may be expressed in miles and in other cases in millionths of a millimeter, but if the unit is fine enough in relation to the accuracy of the measurements, discrepancies will always appear. The fact that repeated sets of measurements never exactly duplicate one another is what is meant by "unreliability." However, at the same time, repeated measurements of a series of objects or individuals will ordinarily show *some* consistency. The block of wood which was the heaviest the first time the set of blocks was weighed will tend to be among the heaviest blocks the second time, and consistency will be the rule among all the blocks of the set. The same, to a degree, will be the case for the weights of the boys in a classroom, or for their performance upon a test of reading comprehension. This tendency toward consistency from one set of measurements to another is the reverse of the fact of variation which we have just considered, and will be designated "reliability."

The consistency of a set of measurements may be approached from two rather different viewpoints. In the first, one is concerned with the actual magnitude of errors of measurement, expressed in the same units in which individual scores are expressed. One thinks of a series of repeated measurements of some characteristic of a particular object, and of the distribution of scores which would result from this repeated measurement. Thus, if a chemical analysis were carried out on successive samples from

a batch of steel in order to determine the percentage of carbon in the steel, the percentage would vary somewhat from sample to sample. If the analysis were repeated 100 times, there would result 100 estimates of the true percentage. These estimates would fall into a frequency distribution, for which measures of central tendency and variability could be computed. The variability of the values in the frequency distribution of repeated measurements, typically expressed as the standard deviation of the distribution, provides a statement of the actual size of the errors of measurement. This statistic is called the "standard error of measurement."

A similar series of repeated measurements could be obtained for anatomical measures such as height or weight. Theoretically, such a series could also be obtained for measurements of reading comprehension, number facility, or any other behavior function. In practice, however, a series of repetitions of the *same* measurement is almost certain to be impossible in the case of human behavior because the individual does not remain the same under the impact of repeated measurements. In educational and psychological testing, the standard error of measurement must always be estimated indirectly by other methods.

A second approach to consistency in measurement may be made in terms of the consistency with which the individual maintains his position in the total group on repetition of a measurement procedure. If two equivalent measures are obtained for each individual within a group, a more or less direct index of the consistency of the measurements is available in the correlation between the two sets of scores. This may be called the "reliability coefficient." For many purposes, the reliability coefficient lends itself to direct and simple interpretation, since it gives directly the proportion of the variance (s^2) of any test score distribution that may be attributed to systematic differences between individuals and not to chance errors. The virtues of the two approaches to the concept of reliability will be compared later in the chapter.

It should already be clear from the discussion that wherever there is reliability in a set of measurements there is also some degree of unreliability. The two are cut from the same pie, being always present together, the one becoming more as the other becomes less. That is, in one situation the consistency of the measurements from one repetition to another may be very marked and the variations quite minor. This would tend to be the case in simple measurements of the common physical properties of a group of objects. In another situation, the consistency may be almost vanishingly small, so there is practically no relationship between an individual's or object's standing in the group upon one set of measurements and the stand-

ing upon another. Theoretically, either consistency or inconsistency may be thought of as approaching zero as a limit, but in practice both are usually present to some degree in any measurement procedure.

The degree of reliability of a set of measurements is a very important consideration, both in the practical day-to-day use of tests and in research projects of various kinds. Some consideration will be given to the importance of reliability, and of data concerning reliability of a particular measuring instrument, in each of these two contexts.

In practical work in measurement and evaluation, we obtain a score for an individual upon some test in order to arrive at some judgment about him, and usually to take some practical action with regard to him. For example, a boy is given a reading test to determine whether he is making satisfactory progress or whether he needs special attention and possibly remedial work. Sometime later he may take a series of achievement examinations so that a college may decide whether he is to be admitted to pursue a course of studies there. Still later, he may be given an interest inventory in order to provide some suggestion as to whether he should be encouraged to specialize in law, medicine, engineering, or some other field. In selecting a test to be used for a practical testing project, and in interpreting the test results, the educator or guidance worker is concerned about the accuracy of the instrument. Other things being equal (in particular, validity) he will always choose the most reliable test from among those available, the one which will provide the most precise estimate of the quality being studied. In interpreting the results from administration of a test, it is always desirable to know how much the obtained score is likely to vary from a true evaluation of the individual's ability to perform on the type of test which has been used.

Clearly, any degree of unreliability in the score resulting from the application of a measuring device is distressing to the educator, guidance worker, industrial personnel officer, or other individual who must use that score as a basis for a practical decision. Unreliability introduces a question mark after the score, and means that any judgment based upon it must be tentative. The lower the reliability of the score, the more tentative the judgment or decision must be, until in the extreme case, as the reliability approaches zero, the score provides absolutely no basis at all for any judgment or decision. The question of the relevancy of the score to the action judgment, though crucial, is a different question and one which falls outside the scope of this chapter. That problem is discussed in chapter 16 on "Validity." The point which is being made here is that as the reliability of a score decreases, the low reliability makes tentative *any* judgment which

is based on that score, and that as the reliability approaches zero, basing *any* judgment on it becomes impossible. The problem of interpretation of scores at different levels of reliability will be discussed more fully later in the chapter.

Reliability becomes of critical importance in research studies at a number of points. In any study of prediction and in any study of improvement resulting from training, *some* degree of reliability in the measure of the criterion being predicted or in the ability being trained is imperative if one is to achieve any prediction on the one hand or any evidence of improvement on the other. One can make no worth-while prediction of a completely unreliable criterion, and one can produce no improvement in a measure of performance which depends entirely upon chance factors. The accuracy of prediction which it is possible to achieve or the amount of improvement in performance which can be shown is limited by the reliability of the measure through which the performance is manifested. Data on reliability of both test and criterion are necessary if the research worker is to be able to interpret the extent to which the imperfect correlation between test and criterion is due to lack of overlapping in function and the extent to which it is due to lack of precision in both measures.

In the analytical study of the relationships among groups of tests information concerning reliability is again crucial. Only with that information available is it possible to determine the extent to which lack of correlation among tests arises because the measures cover unrelated aspects of behavior and the extent to which lack of correlation is due to a lack of consistency within each one of the separate measures.

Logical Considerations in Evaluating Reliability

The evaluation of the reliability of a measuring instrument involves two types of operations, one experimental and the other statistical. On the one hand, it is necessary to apply the instrument to a defined group of cases following a specified experimental design and under specified experimental conditions. On the other, the scores resulting from such administration must be analyzed by appropriate procedures to yield a statistical value which will represent the reliability characteristics of the test. These two aspects are to some extent independent, in that the same essential statistical procedures may be applied to data gathered in quite a variety of ways.

Traditionally, in discussions of reliability determination, the lion's share of the discussion has been devoted to the statistical techniques involved. It is the conviction of the author that much more attenion than has usually been accorded it needs to be given to the experimental aspect. The experi-

mental procedures are very closely bound up with the logical aspects of the problem, so that one must first make an analysis of what is to be accomplished by and what purposes are served by a measure of reliability. The experimental operations must be planned with these purposes in view and evaluated in the light of them. For that reason, the next sections of this chapter are devoted to an analysis of the logical and experimental aspects of reliability. Consideration of statistical procedures follows discussion of the various experimental procedures, a given procedure being discussed in connection with the experimental procedure with which it has the closest connection.

Reliability and Analysis of Variance

Whenever a measuring device is applied to a group of individuals and a score is obtained for each individual in the group, the resulting distribution of scores will spread out over an appreciable range of score values. The variation in any set of scores arises from a number of different factors. Consider measurements of weights of each of the children in a particular school classroom. These differ due to variations in the age of the children, their sex, their parentage, the nourishment which they have received during the years of their life, whether they have been sick recently, whether or not they took a drink of water just before coming to be weighed, the exact angle from which the nurse happened to be looking at the scales, and a host of other factors, minor and major, fugitive and lasting.

The variation in a set of scores arises in part because of systematic differences among the individuals in the group with respect to the quality being measured. In part it arises from unpredictable inaccuracies in the measurement of the separate individuals. Thus, in the example above, variations among children with respect to how large a breakfast they had eaten, how much clothing they had on, how recently they had taken a drink, the angle from which the nurse read the scales, and the like, could be thought of as variable inaccuracies or errors of measurement for different children. These variations would account for some part, though possibly a small one, of the variations in weight recorded from the different children in the class. The evaluation of the reliability of any measure reduces to a determination of how much of the variation in the set of scores is due to systematic differences among the individuals in the group and how much to inaccuracies in measurement of the particular individuals.

There are a number of different statistics which have been developed as summary values to describe the variability in a set of scores. These include the range, interquartile range, average deviation, standard deviation,

and variance. For the purposes of the present discussion, the most useful statistic for describing the variability of a set of scores seems to be the variance (σ^2)[1] or the square of the standard deviation. The particular advantage of the variance, for the present discussion, is that it can be broken down into the separate parts which combine additively to give the total. Thus, if the variance of weight in pounds of pupils in a class were 150, this might break up into a variance of 125 permanently associated with the individuals and a variance of 25 associated with the accidents of that particular set of measurements. These parts add together to make up the total variance of 150 for the set of scores. Whenever a number of independent factors combine to produce a score, it is possible to make an analysis of variance into fractions which are associated with particular factors, and these fractions will sum up to give the total variance.[2] That is,

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 + \cdots + \sigma_k^2$$

where σ^2 is the total variance of the distribution of scores and $\sigma_a^2, \sigma_b^2, \ldots \sigma_k^2$ are the parts of the variance associated with factors $a, b, \ldots, k$ respectively. Thus, the variance in weight of pupils in our classroom might be broken down into variance associated with age, variance associated with sex, variance associated with family, and other variances associated with every other definable stable characteristic of the individuals in the group. There will also be variance which is associated only with the one particular set of measurements, that is, which will not be reproduced another time. This may be designated *"error"* variance. The existence of this error variance corresponds to the fact of unreliability, and its amount relative to the total of all variance is a measure of the degree of unreliability.

It is well to pause briefly at this point and see with just what general type of error we are concerned here when we talk of error variance. Not every type of error, not every discrepancy from the value which an omniscient recording angel would register for the specimen in question qualifies as a part of the error variance. Suppose we were weighing children on scales which were adjusted incorrectly in such a way that on the average each child was given a weight two pounds above his "true" weight. This adjustment error is uniform and systematic and results in a "constant error." As described, it actually does not contribute to variance at all, though it does make every observation incorrect. In our present discussion, we are

[1] In symbolism for representing the variance of a distribution, σ^2 will be used to represent the theoretical population value, while s^2 will designate the value obtained from a specific limited sample.

[2] When factors are not independent, it becomes necessary to analyze covariances as well as variances.

not concerned with such constant errors, disastrous though they may be to scientific precision. Again, suppose that the scales on which we were weighing children were in error in such a way that they credited each child with two pounds for each pound he weighed in excess of 50 pounds. This would not be a *constant* error, but it would be systematic, that is, it is statable in definite terms and predictable for any child. This is still not the type of error with which we are concerned in discussions of reliability. The type of errors which we have in mind when we speak of "chance errors of measurement" are errors which are esentially unpredictable from anything we may know about the individual or his previous performance. These errors may be defined as follows:

$$\sum e_1x_1 = \sum e_2x_2 = \sum e_1e_2 = 0$$

where e_1 and e_2 refer to error on two forms of a test, and x_1 and x_2 each refer to "true score"[3] on the corresponding form. What we are saying is that the type of error in which we are concerned is error which is unrelated to the individual's true score or to his error on another form of the test.

Let us designate the variance of true scores of a group on a trait by $\sigma_\infty{}^2$ and the variance of errors of measurement by $\sigma_e{}^2$. If the magnitude of the error of measurement is unrelated to the magnitude of the true score, we have,

$$\sigma^2 = \sigma_\infty{}^2 + \sigma_e{}^2.$$

That is, the variance of the obtained scores equals the sum of the variance in true scores and the variance arising from errors of measurement. It is also possible to relate these fractions of variance to the reliability coefficient which was discussed earlier in the chapter. We have

$$r_{11} = \frac{\sigma_\infty{}^2}{\sigma^2} \tag{1}$$

and

$$r_{11} = 1 - \frac{\sigma_e{}^2}{\sigma^2}. \tag{2}$$

That is, the numerical value of the reliability coefficient of a test corresponds exactly to the proportion of the variance in test scores which

[3] We shall repeatedly have occasion to use the expression "true score" throughout this chapter. The term is convenient but a little misleading. As we speak of it, true score is not the ultimate fact in the book of the recording angel. Rather, it is *the score resulting from all determinable systematic factors,* including any systematic biasing factors which may produce systematic incorrectness in the scores. This larger expression should be understood whenever the term "true score" is used.

is due to true differences between individuals in the quality being evaluated by the test. A test is unreliable in proportion as it has error variance.

It becomes clear that the basic problem in determining the reliability of a testing procedure becomes that of *defining* what shall be thought of as true variance between individuals and what shall be thought of as error variance. When this definition has been reached, the next step is to devise those series of experimental and statistical operations which will provide the best estimates of the defined fractions of variance. The next section will deal, therefore, with the analysis of types and sources of variance in test scores. After this analysis, various experimental operations which have been proposed to provide data for estimating reliability will be considered, each set of operations being evaluated in terms of the logical analysis.

Sources of Variance in Test Scores

As noted above, variance in a set of scores from any test or measuring device arises from a great variety of specific sources. However, these may profitably be grouped, for purposes of discussion, into a few major categories. A classification of sources of variance is presented in Table 8.[4] The categories given here probably do not exhaust the possible range of categories. Certainly, many more subcategories could be listed under most of the major headings, and those which are presented should be thought of as illustrative rather than exhaustive. A consideration of each of the categories will provide the basis for a decision as to which fractions of variance should be thought of as true, systematic variance in the quality or qualities being measured and which should be thought of as error variance.

Variance within a set of scores arises first of all because different individuals possess different amounts of certain general and persistent traits (category I of Table 8). Thus, in a series of intellectual tests some type of ability to reason deductively might be a general quality which entered into a number of the tests and which, for each of the tests, accounted for part of the individual differences in performance. Or several arithmetic tests might have a common factor of facility with numbers. Verbal comprehension is likely to enter into a wide range of tests requiring reading. Almost any test performance will depend in part upon general abilities which are also involved in a number of other types of test performance. The type of variance which is now under discussion represents a persistent, lasting characteristic of each individual, causing stable individual differences in test performance. Since it arises from a persisting feature of each individual, this variance is clearly systematic variance and should be so treated in any sequence of operations set up to provide an estimate of reliability.

[4] A similar analysis, differing in detail, has been formulated by Cronbach (*3*).

TABLE 8

Possible Sources of Variance in Score on a Particular Test

I. *Lasting and general characteristics of the individual*
   A. Level of ability on one or more general traits, which operate in a number of tests
   B. General skills and techniques of taking tests
   C. General ability to comprehend instructions

II. *Lasting but specific characteristics of the individual*
   A. Specific to the test as a whole (and to parallel forms of it)
      1. Individual level of ability on traits required in this test but not in others
      2. Knowledges and skills specific to particular form of test items
   B. Specific to particular test items
      1. The "chance" element determining whether the individual does or does not know a particular fact. (Sampling variance in a finite number of items, not the probability of his guessing the answer.)

III. *Temporary but general characteristics of the individual*
   (Factors affecting performance on many or all tests at a particular time)
   A. Health
   B. Fatigue
   C. Motivation
   D. Emotional strain
   E. General test-wiseness (partly lasting)
   F. Understanding of mechanics of testing
   G. External conditions of heat, light, ventilation, etc.

IV. *Temporary and specific characteristics of the individual*
   A. Specific to a test as a whole
      1. Comprehension of the specific test task (insofar as this is distinct from I B)
      2. Specific tricks or techniques of dealing with the particular test materials (insofar as distinct from II A 2)
      3. Level of practice on the specific skills involved (especially in psychomotor tests)
      4. Momentary "set" for a particular test
   B. Specific to particular test items
      1. Fluctuations and idiosyncrasies of human memory
      2. Unpredictable fluctuations in attention or accuracy, superimposed upon the general level of performance characteristic of the individual

V. *Systematic or chance factors affecting the administration of the test or the appraisal of test performance*
   A. Conditions of testing—adherence to time limits, freedom from distractions, clarity of instructions, etc.
   B. Unreliability or bias in subjective rating of traits or performances

VI. *Variance not otherwise accounted for* (*chance*)
   A. "Luck" in selection of answers by "guessing"

Two rather special types of persisting general factors deserve some particular mention. These are the general ability to comprehend instructions and what we may speak of as "test-wiseness." These factors are mentioned

because they are likely to enter into any test score, whether we want them to or not. That is, performance on many types of tests is likely to be in some measure a function of the individual's ability to understand what he is supposed to do on the test. Particularly as the test situation is novel or the instructions complex, this factor is likely to enter in. At the same time, test score is likely to be in some measure a function of the extent to which the individual is at home with tests and has a certain amount of sagacity with regard to tricks of taking them. Freedom from emotional tension, shrewdness with regard to when to guess, and a keen eye for secondary and extraneous cues are likely to be useful in a wide range of tests, particularly those which are not well constructed. The presence of variance in score due to variation in comprehension of instructions and in test-wiseness is usually undesirable from the point of view of the purposes of the test in question. It usually represents systematic invalid variance serving systematically to reduce the validity of the test. However, these factors must be recognized. They present a challenge to the author of the test, who will try to minimize them, except where their presence is specifically desired, by providing the clearest possible instructions and a minimum of secondary cues. These factors present a problem of validity rather than one of reliability; as far as our present analysis is concerned, they represent a general, lasting quality of the individual and must be treated as such.

In addition to variance which is common to a range of tests, each test will have some variance which arises from persistent characteristics of the individuals being studied but which is specific to the particular area being tested (category II). That is, there is some variance which will be present in spelling tests, for example, but not in tests of any other performance. There are, of course, degrees of specificity of knowledge or skill, so that further narrowing down may take place even within a given field. In addition to variance which characterizes the field of performance, such as spelling or numerical computation, there may be variance associated with the specific form and manner of testing. In the case of spelling this might relate to oral presentation as in a spelling bee, writing words from dictation, or recognition of errors in words presented in a printed test. A numerical operations test might be presented in free-response or multiple-choice form, and variance might be associated with that feature. Finally, in any test there is likely to be variance associated with the particular sample of test items. There will be a certain amount of variation in specific bits of knowledge or skill, so that even the individual who has high over-all ability in the area in question will lack certain specific items of knowledge or skill and the individual low in general performance will succeed on isolated items not

known by his generally more proficient fellow. The sampling of items—words to be comprehended, formulas to be known, generalizations to be applied, and the like—will be a source of variation in the resulting test scores. Given two tests made up of samples independently chosen in the same way from the same universe of items, individuals will fail to receive identical scores on the two tests because of variation in the particular items which each individual happens to have the skill or information to answer.

At this point we begin to encounter some difficulty in the logical determination of what shall be allocated to systematic variance and what to error variance. Variance specific to the area covered by the test (category II A 1) is certainly systematic variance, and any operation for determining reliability should be so planned that this type of variance is treated as systematic. The problems arise in connection with variance associated with the particular test format and with the particular sampling of test items. The question is one of finding the most useful definition. How broadly shall we define what we are measuring? Shall we define it in terms of an area of content only? In terms of an area of content and form of test? In terms of an area of content, form of test, and particular set of test items?

The first definition above leads to experimental procedures which appear to come closer to evaluating test validity than test reliability. That is, if format is considered a source of error variance, we are led to correlate tests with different types of items and manner of presentation. We are then beginning to inquire whether the test is consistent with other measures rather than whether that particular test measures consistently. The third definition appears so narrow as to have little practical meaning in most cases. We are rarely interested in performance on a limited set of test items for their own sake. We are almost always interested in test performance as an indication of ability to perform on the whole universe of items of which the test represents a limited sample. The variance due to the sample of items is, therefore, in an entirely true sense part of the error of measurement. In conclusion, then, it seems that the most meaningful definition of reliability will allocate to systematic variance that variation arising from the specific abilities required in the area being studied (category II A 1) and that arising from knowledges and skills specific to the particular form of the test (category II A 2), but will allocate to error variance that variation in performance arising from the particular sampling of test items (category II B).

The above discussion brings home one very important point. There is no single, universal, and absolute reliability coefficient for a test. Determination of reliability is as much a logical as a statistical problem. The appropriate

allocation of variance from different sources calls for practical judgment of what use is to be made of the resulting statistical value. This point will become increasingly apparent as the discussion continues.

A third group of factors making for variation in test scores are certain general but temporary characteristics of the individual or of the testing situation. These include such factors as state of health, amount of sleep the previous night, presence or absence of worries or other distracting influences, and a host of other internal factors which may have bearing upon the efficiency of the individual's work. Different test performances will be susceptible to these factors in varying degrees, but all will probably be influenced by them in some measure. The factors vary both in their permanence and their generality. Some may change from day to day, some from hour to hour. There may even be very short time fluctuations in efficiency which represent a change from minute to minute. In general, however, we may think of these factors as ones which characterize an individual at a particular testing session but not at another session.

Here, again, a problem arises as to what allocation is to be made of variance of this type. Once again, the problem becomes that of determining the type of consistency which it seems significant to measure. Is it important to determine how consistent a measure we have of the individual as he exists at a particular moment? Or is it important to determine how consistent his performance is from day to day and week to week? For some purposes the former may be the significant information, for some purposes the latter. If our interest lies in studying the intercorrelations among a battery of tests which have been given at one time, the appropriate measure of reliability for use in conjunction with those correlations would seem to be a measure of consistency at that moment in time. However, if the test results are to be used for predicting something about the individual at some later date or evaluating the result of training over some extended period, the more meaningful definition of reliability would appear to be that phrased in terms of consistency over a period of time. There are other specific purposes for which tests might be used, and in each case it will be necessary to decide whether it is more meaningful for the temporary characteristics of the individual (category III) to be thought of as a source of systematic variance or as a source of error variance.

Our general discussion so far has provided no indication as to whether this or any other *possible* source of variance does in fact yield practically significant amounts of variance. That is, we have not shown whether or not it makes any *practical* difference what we do with variance in the above category. That cannot be a matter of general theoretical discussion, but

must be a matter of specific empirical evidence in each case. The answer will probably vary widely in different areas of measurement. We might guess, for example, that in a simple power test of vocabulary less of the variance would be accounted for by temporary characteristics of the individual, than in a test of mood or feeling tone, for which substantial day-to-day swings might be expected.

A further group of factors making for variation in test performance is made up of certain relatively temporary and specific factors. In this category are included influences which tend to be more limited both in time and in scope than those discussed in the immediately preceding paragraphs. Certain of these factors characterize performance on a test as a whole. If the test is novel and the instructions difficult, individuals may vary in the extent to which they "catch on" to the nature of the task. In part this will probably represent general ability to understand instructions (category I B), but in part it may represent temporary or "chance" variations superimposed on that general ability. Again, a test may call for certain specific tricks or techniques of which the individual does or does not "get the hang." Furthermore, performance on many tests, particularly measures of complex coordination or skill, is susceptible to considerable improvement through practice. A temporary feature of some importance may be the individual's practice level at the moment of testing. Finally, there are certain factors which, for the lack of any better term, we may group together under the heading of "mental-set" at the time of taking the test. Was the subject emphasizing speed or accuracy if it was a speeded test? To what cues was he particularly alert if it was a perceptual task? What was his momentary mood if it was an attitude or interest questionnaire?

The factors grouped in this category (IV A) are the ones whose presence and significance in any given case are probably most open to question. In many types of simple and standard tests they can perhaps be ignored. However, in novel types of tests, highly speeded tests, measures into which introspective interpretation enters heavily, and perhaps other types, the possibility of encountering variance from such sources as have just been discussed must be given serious consideration. The rationale for allocation of this type of variance would appear to follow much the same lines as that for variance attributable to general temporary conditions (category III) discussed previously.

In addition to factors specific to a particular test and date of testing, there may be even more specific and temporary factors. These are factors which are specific to an item or a few items and a minute or a few minutes of time. These factors appear as short-time fluctuations of memory or attention, mo-

mentary blockings of performance, cyclic variations in effort, and a variety of other fluctuations superimposed upon the general level of performance. These factors (category IV B), insofar as they affect score, introduce variable and unpredictable error into the score, and the treatment of the resulting variance should always be such that it is allocated to error variance.

For some situations we must recognize sources of variance not only in the subject being tested but in the conditions of giving and appraising the test. On the one hand factors of timing of a test, test instructions, amount of noise and distraction, and the like may vary from person to person or subgroup to subgroup. One can see that this type of variation is especially likely to arise in the case of tests which are individually administered by a number of examiners, tests which are closely timed, or tests which have very complex instructions. On the other hand, variance may be introduced in appraising the test performance or other behavior which is to yield a score. This is true in proportion as the appraisal depends on the judgment of another human being. In essay examinations, rating scales of all sorts, projective tests, or in fact anything which calls for interpretation or evaluation by an observer or scorer, variance due to the scorer will enter in. There will be variance associated with different scorers, variance associated with changes in the scorer from time to time, and variance representing unpredictable scorer inconsistency. These types of variance (category V) become important only in certain particular measurement situations; but wherever they occur they constitute error variance, and procedures should be planned to identify them and treat them as error.

Finally, we must introduce the concept of "chance" to take care of variance not otherwise accounted for. We can never find antecedent factors which will account for all the variance in a set of test scores. Some variance arises from guessing at answers, some from other obscure variable influences which we cannot define or specify. The variance of this type (category VI) is error variance in its purest form, and the operations which define reliability must allocate it to that category.

Procedure for Estimating Reliability

The evaluation of the reliability of a measuring instrument requires a determination of the consistency of repeated measurements of the same object or group of objects. In the physical sciences many repetitions of a measurement of a single object or phenomenon may provide a reasonable method for estimating the precision of the measurement procedure. In dealing with human behavior, however, the individual is likely to be changed as a result of the operation of measurement, and it usually is neces-

sary to limit sharply the number of times a single individual is measured. In practice, therefore, all procedures of reliability estimation generally useful to psychology and education are based upon getting a small number of measurements, typically only two, for each individual in a representative group. Stability of results is achieved by increasing the number of individuals measured rather than the number of measurements of each. These measurements provide sets of scores, again usually two for each individual, for analysis. The usual analysis has consisted of computation of the coefficient of correlation between the two sets of scores, yielding an estimate of average consistency for the group.

We have defined the reliability coefficient of a test as the ratio of the variance of true scores to the variance of obtained scores (made up of true measure and error). Where two equivalent forms of a test are available, it can be shown that this reliability coefficient is in fact the correlation between the two forms. Equivalent forms in this situation are defined as tests which overlap completely in their true score variance, and for which the proportion of true score to total variance is the same.

We must now determine how equivalent measures may be set up, so that the correlation between them may be obtained, and how the true variance of a set of scores may be estimated so that its ratio to the total variance may be calculated. These are, in fact, one and the same problem. Equivalent tests will be defined as tests which have identical true variance, but no overlap in error variance. Or the reverse, the true variance of a set of test scores will be defined as that which is common to that test and an equivalent test. We must next consider what actual testing operations will correspond satisfactorily to the logical requirements for equivalent tests. These logical requirements have been considered in the previous section, in connection with the desired allocation of different fractions of variance.

A number of different testing and statistical procedures have been proposed to provide the necessary coefficient of correlation between equivalent measures. Many of these represent efforts to develop short-cuts to the preparation and administration of two separate tests built to the same set of specifications, and therefore assumed to be equivalent. Others have been defended as preferable procedures. We shall consider the major sets of experimental and statistical procedures in turn, describing each and evaluating each in terms of its treatment of different categories of variance. The major procedures are:

1. Administration of two equivalent tests and correlation of the resulting scores.

2. Repeated administration of the same test form or testing procedure and correlation of the resulting scores.

3. Subdivision of a single test into two presumably equivalent groups of items, each scored separately, and correlation of the resulting two scores.

4. Analysis of the variance among individual items, and determination of the error variance therefrom.

A section will be devoted to each of the above procedures in turn, describing various subprocedures and discussing the problems which arise with each.

1. Reliability Defined by Equivalent Test Forms

Since the formal definition of reliability has been phrased in terms of the correlation between two equivalent sets of measures, it seems obvious that the procedure for reliability determination which makes use of two equivalent tests will measure up to our logical requirements. This is in fact true, provided we can establish satisfactory procedures for preparing truly equivalent tests. This is a problem in the logic and practice of test construction. In preparing equivalent test forms, there is danger, on the one hand, that the two tests will vary so much in content and format that each will have some specific variance (category II A) distinct from the other, in which case the correlation between the two will underestimate the reliability. There is the reverse danger that the two forms may overlap to such an extent in specific details of content that variance due to specific sampling of content (category II B) may be common to the two tests. In that case, this variance will be treated as systematic rather than chance variance, and the obtained correlation will overestimate the reliability.

The best guarantee of equivalence for two test forms would seem to be that a complete and detailed set of specifications for the test be prepared in advance of any final test construction. The set of specifications should indicate item types, difficulty level of items, procedures and standards for item selection and refinement, and distribution of items with regard to the content to be covered, specified in as much detail as seems feasible. If each test form is then built to conform to the outline, while at the same time care is taken to avoid identity or detailed overlapping of content, the two resulting test forms should be truly equivalent. That is, each test must be built to the specifications, but within the limits set by complete specifications each test should present a random sampling of items. In terms of the practical operations of test construction, it will often be efficient to assemble two equivalent test forms from a single pool of items which have been given preliminary tryout. Within the total test if the test is homogeneous in the character of its content, or within parallel homogene-

ous sections of a heterogeneous test, items from the pool should be assigned to the two forms in such a way as to give the same distribution of item difficulties and the same distribution of item-test correlations in each form.[5]

Two tests constructed in the above way will treat as systematic variance that variance in categories I and II A of Table 8. They will treat as error variance that in categories II B, IV B, and VI. The allocation of variance in categories III and IV A will depend upon the time interval between the administration of the two forms. If they are given in immediate sequence, this last variance will be treated as systematic variance; if some time intervenes between the testings, this variance will be allocated to error. For most uses of the resulting statistic, it will probably be more meaningful to let some time elapse between the two testings, thus treating temporary day-to-day fluctuations as errors of measurement. The question arises as to how long an interval should elapse. For most purposes, the answer to this lies in the thought that it is day-to-day fluctuations which we wish to allocate to error. An interval of a few weeks would usually appear sufficient. With longer intervals the problem of genuine growth and change in the individual is encountered, and the coefficient may be lowered because of these changes. Of course, for some purposes we may be interested in consistency of performance over an extended period of time, but consistency of this type represents a rather different concept of reliability.

In most of the usual types of tests of ability or achievement, preparing equivalent forms should not present undue difficulty. There are some situations, however, in which equivalence will be very difficult to achieve. This is true when either (*a*) the test task is essentially unique or (*b*) a single exposure to the test changes the individual to such an extent that he is really a different individual at the second exposure. The former case may occasionally arise in connection with unusual problem-type tasks. The second case is the more common source of difficulty. In any task which is sufficiently novel so that the experience of being tested adds a significant increment to the individual's practice with the task, he is a somewhat different individual at the time of a subsequent test. In novel tasks, or in

[5] If homogeneity is assumed for the materials in a test (or subtest), some check upon the degree to which equivalence has been achieved may be obtained by examining item correlations within and between the two test forms. On the average, these should be equal. An estimate of the extent to which this is the case might be obtained by subdividing each of the forms into two halves, by some such procedure as taking alternate items, and getting all the correlations among the resulting four scores. If we find approximately that

$$r_{12}r_{34}=r_{13}r_{24}=r_{14}r_{23},$$

this evidence supports the equivalence of the two tests.

tasks which present essentially a learning situation, the changes may be quite marked. The problem of defining reliability for such a changing function is a very difficult one, and no completely satisfactory solution seems available.

2. Reliability Defined by Repetition of Identical Test Form

In some cases, obtaining two equivalent measures will reduce to repetition of identically the same measuring instrument, the only difference being the time at which it is administered and perhaps the person by whom it is administered. That is, two equivalent measures of weight could be obtained by weighing the members of the group being studied on the same scales at two times ten days apart. The same situation would hold for almost any physiological or anatomical measurement. In these cases, we do not encounter the problem of sampling items from a larger universe of behavior, and so distinct equivalent test forms are neither meaningful nor possible. Equivalence in this case means identity of measuring instrument and procedure.

There are also certain behavioral measures in which the situation of sampling from a large universe of items does not arise. This is the case in simple repetitive tasks of motor speed and skill or of perceptual judgment. Thus, in a test of simple reaction time, in which a measurement of the individual is obtained by timing repeatedly his simple reaction to some stimulus, the test task is so defined that no varied sampling from a more extensive universe of behavior is involved. Here, again, repetition of the same test provides the meaningful definition of "equivalent measures." The same would tend to be true of the type of perceptual judgment which is involved in the simple psychophysical experiment, as with judgments of brightness, length, weight, and so forth. In all these cases, repetition of the test appears to provide an acceptable procedure for reliability determination.

In most measures of intellect, temperament, or achievement, however, repetition of the same test form and correlation of the two sets of scores is less defensible as an operation for determining reliability. In these cases, a particular test consists of a limited sample from a much larger universe of possible items. The test score has practical significance insofar as it is representative of the individual's ability to respond to all of the tasks in the universe which it undertakes to sample. Reliability is a matter of the adequacy of the sampling of items as well as the consistency of behavior by each individual. In other words, in this case sampling of items (category II B) is an appreciable source of variance. Practical usefulness for the

result dictates that this variance be treated as error variance in determining the reliability of the test. Repeating the same test form holds the sampling of items constant so that this factor is treated as systematic rather than error variance. Reliability coefficients calculated from a repetition of the same test may be expected to be higher than those based upon parallel, equivalent forms to the extent that this variance associated with sampling of items is a factor.

A second possible difficulty with repetition of the same test, which may or may not be important in any given case, is actual memory of particular items and of the previous response to them. Insofar as this memory is effective in leading the individual to repeat the same response he made the time before, the results on two test administrations tend to be abnormally alike. The same answers may be repeated not because the individual is consistent in his behavior, and arrives at the same conclusion in the same way, but because he happens to have a memory of his previous response. In effect, some of the variance associated with momentary memories and chance choices (categories IV B and V) becomes common to the two testings and is treated as systematic variance. Memory of previous responses is likely to be a factor in proportion as (*a*) the test is short, (*b*) the test items are distinctive and memorable, and (*c*) the interval between testings is short.

Another element that should be considered in deciding whether a repeated test is comparable to the original test is the attitude of the person tested. Especially where a test is quite long, as is the case with some interest and personality inventories, for example, repetition may be tedious for the subject, and he may therefore give more haphazard responses. This would, of course, operate to lower the correlation between the two testings. If a test score is likely to be greatly affected by motivation, as when the test requires very rapid or very concentrated work, it is especially important that the subject should feel that both testings will have equal significance for him. If one were to retest a group of men who had been accepted for pilot training, for example, and correlate these scores with scores earned during the initial classification period, before the men were sure of being accepted, the changed motivational conditions might well reduce the correlation and yield an underestimate of reliability. This matter of changed attitude and motivation could also affect a retest with an equivalent form of the test, but is probably likely to be most acute when the same test form is repeated.

In summary, for those types of tests in which sampling of items and memory of previous responses are not an issue and for which reasonable

comparability of motivation seems likely, a second application of the same test at a later date and correlation of the two sets of scores provides an adequate set of operations for reliability estimation. In the many other cases, however, in which the factors of sampling and memory are significant sources of variance, repetition of the same test form will yield an estimate of reliability which tends to be systematically too high. In these latter cases the procedure is to be avoided. The correlation will also be unrepresentative, and probably too low, if the attitude of the subject changes from one testing to the other.

3. Reliability Defined by Subdivision of Single Total Test

The preparation and administration of two equivalent test forms, though quite satisfactory as a procedure for estimating reliability, presents certain practical difficulties. These center around the problems of the time and labor involved both in the construction and the administration of two complete test forms. If only a single form of a test is needed for the research or practical use to which the test is to be put, it often seems unduly burdensome to prepare two separate tests merely in order to obtain an estimate of reliability. Furthermore, when a test is developed and administered as part of a research project, time for the administration of an equivalent form of the test is often not conveniently available. In the interests of economy it becomes desirable to set up procedures for extracting an estimate of reliability from a single administration of a single test. One group of such procedures subdivides the total test artificially into two half-length tests and correlates the scores on those. This correlation gives the reliability, not of the full test, but of one only half as long. The reliability of the full test must be estimated by the use of formula (3) or (4), see page 580. The second group of procedures is based essentially upon the analysis of variance among single test items. The procedures for subdividing the test will be considered in this section, and the next section will be devoted to procedures based upon analysis of the single items.

Manner of assembling part scores

If a test is composed of $2n$ separate items or parts, there are $\frac{(2n)!}{2\,(n)!(n)!}$ ways in which two subtests, each composed of n items, can be assembled from it. Certain procedures have been proposed for selection from among these possible alternatives, either on logical grounds or on grounds of convenience. The more usual procedures include: (*a*) selecting sets of items

for the two half-tests which appear equivalent in content and difficulty, (*b*) putting alternate items or trials in each half-test, (*c*) putting alternate groups of items or trials in each half-test, (*d*) using the first half of the items or trials as one half-test and the second half as the other.

We must consider the specific merits of these different procedures, together with questions as to the logical acceptability of split-test procedures in general. Before entering into these considerations, however, it will be appropriate to indicate how we may obtain an estimate of the reliability of the *whole* test from the correlations between two *half-tests.*

Reliability of total test from part-test correlations

It can readily be shown that the correlation between the sum of two sets of scores, x_1 and x_2, and the sum of two other sets of scores, x_3 and x_4, is given by the formula

$$r_{(x_1+x_2)(x_3+x_4)} = \frac{r_{13}\sigma_1\sigma_3 + r_{14}\sigma_1\sigma_4 + r_{23}\sigma_2\sigma_3 + r_{24}\sigma_2\sigma_4}{\sqrt{\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2}\sqrt{\sigma_3^2 + \sigma_4^2 + 2r_{34}\sigma_3\sigma_4}}.$$

Let x_1 and x_2 represent scores on two halves of one form of a test and x_3 and x_4 represent scores on two halves of another equivalent form of the test. If we now assume that all terms of the type $r_{ij}\sigma_i\sigma_j$ are equal, so that any one of them can be represented by $r_{12}\sigma_1\sigma_2$, and that the standard deviations of the two full-length tests are equal, we get

$$r_{(1+2)} = \frac{4r_{12}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{12}}, \tag{3}$$

where $r_{(1+2)}$ is the reliability of the full-length test made up of 1 and 2. If it is assumed further that $\sigma_1 = \sigma_2$, this expression simplifies to

$$r_{(1+2)} = \frac{2r_{12}}{1 + r_{12}}. \tag{4}$$

This is the familiar Spearman-Brown formula, originally presented by Spearman (*23*) in 1910, for estimating the reliability of a measure from the score on a smaller segment of behavior.

When the standard deviations of the two half-length tests are actually equal, formulas (3) and (4) will give identical results. When the two standard deviations are not actually equal, formula (4) will give a higher value as the estimate of $r_{(1+2)}$. This higher value may, in fact, be the better estimate, since the assumption that the terms $r_{ij}\sigma_i\sigma_j$ are equal may not be a very defensible one if the σ's are not equal. In practice, differences between the results by the two formulas are likely to be small.

A number of expressions have been developed which are algebraic equivalents of (3), but which are based on different ones of an interrelated set of values. A number of equivalent forms of formula (3) are presented in Table 9. This table indicates the values which are required by each formula, the formula itself, and the source for the formula.

TABLE 9

EQUIVALENT FORMULAS FOR ESTIMATING RELIABILITY FROM HALF-LENGTH TESTS

Entering Statistics	Formula	Source
$r_{12}\ \sigma_1\ \sigma_2$	$\dfrac{4\sigma_1 \sigma_2 r_{12}}{\sigma_1^2 + \sigma_2^2 + 2\sigma_1 \sigma_2 r_{12}}$	Flanagan in (*21*)
$\sigma_1\ \sigma_2\ \sigma_t$	$2\left[1 - \dfrac{\sigma_1^2 + \sigma_2^2}{\sigma_t^2}\right]$	Guttman (*8*)
$\sigma_1\ \sigma_t\ r_{1t}$	$\dfrac{4(\sigma_1 \sigma_t r_{1t} - \sigma_1^2)}{\sigma_t^2}$	Mosier (*16*)
$\sigma_{(1-2)}\ \sigma_t$	$1 - \dfrac{\sigma_{(1-2)}^2}{\sigma_t^2}$	Rulon (*21*)
$\sigma_1\ \sigma_{(1-2)}\ r_{1(1-2)}$	$\dfrac{4(\sigma_1^2 - \sigma_1 \sigma_{(1-2)} r_{1(1-2)})}{4\sigma_1^2 + \sigma_{(1-2)}^2 - 4\sigma_1 \sigma_{1-2} r_{1(1-2)}}$	Cronbach*

* This form of presentation, together with the last formula, was suggested in a personal communication by Lee Cronbach.

Formula (4) may be generalized to any increase in the length of the test, and it then becomes

$$r_{nn} = \frac{nr_{11}}{1 + (n - 1)r_{11}}, \tag{5}$$

where r_{nn} is the reliability of a test n times the length of the test from which the observed correlation, r_{11}, was obtained.

The assumptions which were made in arriving at formula (5) should be noted. They were equality of the standard deviations of the part scores and equality of the part score intercorrelations. These conditions will be satisfied in those cases in which (*a*) the part scores are equivalent, in the sense described on page 575 (that is, the specifications for each part are the same in terms of content, number of items, distribution of item difficulties, and distribution of item internal consistency measures), and (*b*) the nature of the function measured by later items or trials of a lengthened test is not changed as a result of the experience with earlier items or trials. However, Kelley (*13*) has indicated that formula (5) is not sensitive to

differences in variability between the fractions of the test or even to differences in the level of reliability of the sections of the test. The value given by the formula is a close approximation to the actual correlation of longer tests even when the standard deviations and reliabilities of the unit-length tests show appreciable variation. The more serious limitations lie in the manner in which the experimental procedures allocate certain fractions of the variance. These points will be discussed presently.

General evaluation of reliability estimation from part scores

In general, estimating reliability from two parts of the same total test differs from reliability estimation from the administration of two separate tests in two respects: (*a*) the two parts are not separately timed, and (*b*) the performances on the two parts are necessarily adjacent or even intermingled in time. A question may be raised as to the comparability of the two part scores, but the same issue arises with regard to any two test scores, whether they stem from artificial subdivisions of the same total test or from separate and distinct test forms.

The use of a single common time limit for a test becomes of critical importance whenever the test is in some degree a speed test. This can be seen most clearly by considering a pure speed test, in which each individual could do each item if he were given enough time and in which individuals differ only in the number of items which they can do within the limited amount of time available. In this case it is completely impossible to extract two meaningful scores from a test with a single time limit. The score which an individual makes upon a group of items will depend solely upon where the items are placed in the test. If the two part scores are made up one of the odd-numbered and the other of the even-numbered items, each individual will *necessarily* have practically identical scores on the two halves, because opportunity to attempt items has been systematically equated for the two half-scores. On the other hand, if one half-score is made up of the first half of the items on the test and the other of the second half, scores on the two halves cannot possibly be compared meaningfully because the individual can score on the second part only insofar as he has already completed and got a perfect score on the first part.

In practice, no test is an absolutely pure speed test of the sort which we have imagined in the previous paragraph. However, there are a good many tests which involve speed to some extent. Insofar as speed, as distinct from "power," or level of performance with unlimited time, is a factor in the test performances of a group of individuals, the results from split-test procedures for determining reliability will lack meaning. The amount

of distortion of the results will be a function of the extent to which individual differences in score depend upon individual differences in speed of performance. In general, split-test estimates for speeded tests are misleadingly high. This fact has not always been appreciated by test authors and publishers, and in interpreting published reliability data the reader must watch out for this misuse of the split-test method.

The second limitation upon split-test reliabilities is the lack of time interval between the two performances. The two performances are not only adjacent in time, but in most cases even intermingled. This means that the day-to-day fluctuations in conditions (category III) and even the minute-to-minute variance in performance (category IV) are equated for both part scores and tend to be allocated to systematic rather than error variance. This means that split-test reliabilities, even of wholly unspeeded tests, may be expected to be in error on the high side, and insofar as variance of the above types is substantial we may expect a substantial overestimation of the effective reliability of the test.

We shall now turn our attention to the various possible ways of subdividing a test, and consider the specific limitations and advantages of each.

Selection of half-tests equated for content and difficulty

Since the total test was presumably constructed to conform to certain specifications (as to content, difficulty, and the like), it is only reasonable that each of the half-tests should conform to these same specifications. The best guarantee of equivalence in the two half-tests would seem to be that the items for each be specifically chosen so as to be equivalent. In other words, the same procedures which were described on page 575 as appropriate in the construction of the two equivalent forms can be applied to the problem of subdividing the items within a single form. Items should be selected for each half-test so as to make it conform to the specifications for the total test, but within these limits chance should determine which items go in which half of the test.

This would appear to be the most defensible procedure for obtaining two half-scores from a test. Its disadvantages are chiefly practical ones, in that it requires work, and also ideally some data about the individual items, to obtain effective equivalence. It may also involve some sacrifice of convenience in scoring the tests.

A compromise between the equivalent-forms and split-test procedures should be noted, at this point, which appears to gain many of the advantages of both procedures. This is to prepare only the amount of testing

material which is to be used in the final testing operation, but to arrange it in two equivalent and separately timed halves. If convenience requires it, the two half-length tests can be administered in immediate succession, but with separate time limits. Except that the two scores are obtained from segments of behavior which are immediately adjacent in time, this procedure is the counterpart of testing with separate comparable forms, yet only enough test materials for a single test need be prepared. In other cases, the two half-length tests can be given on different days. If that preliminary tryout of the test materials which should be expected of any good testing instrument is carried out, preparation of suitably equivalent half-tests should not be difficult. This permits the determination of an equivalent-forms reliability without the labor incident to constructing two full-length tests. The separate timings of the halves will permit studies of reliability for research purposes. The sum of the two half-test scores will, of course, provide a total score for the test. If the test is to be widely used for other than research purposes, in the edition which is prepared for general use, it may be desirable and will usually be quite practical to merge the two halves into the more usual test with single timing. This procedure is strongly recommended to research workers.

Alternate items as a basis for splitting test

A procedure which has been widely adopted for splitting a test to yield two half-scores is that of treating the odd-numbered items as one half-test and the even-numbered items as the other. This procedure has simplicity and objectivity as considerations to recommend it. It also appears to be related to the frequent practice of grouping together in a test form items of similar structure and of graduating the difficulty of the items from easy to hard. If the items within the test are arranged in this systematic fashion, the odd-even procedure provides a simple way of approximating equivalence in the two half-scores. If there are several successive items on the same topic or of the same type, this procedure will automatically divide them evenly between the two half-tests. If the items progress in difficulty, approximate equivalence of the half-tests in difficulty level is guaranteed. However, this makes very definite assumptions, which may not be warranted in a given case, as to the manner in which the test items are arranged in the full test.

The odd-even items split provides, at best, only a rough-and-ready approximation to equivalence in the two half-tests. It is particularly in the case of this procedure that certain other issues are raised, which we may well discuss now. If, either because of very close similarity of content or

because of moment-to-moment fluctuations in efficiency, performance on successive items tends to be more alike than upon items widely separated in the test, we may find that the error of measurement in successive items is not independent but correlated. That is, variance in categories II B and IV B of Table 8 which should be treated as error variance may be common to successive items. By systematically allocating successive items to the two half-scores, this variance becomes common to the two and is treated as systematic variance. The odd-even procedure is pre-eminently the one which makes short-time fluctuations in individual performance operate to give an appearance of reliability rather than the reverse.

Alternate groups of items as basis for splitting test

In some cases, several items may be unduly closely related in content. Examples of this would be a group of reading comprehension items all based on the same passage, a group of items all referring to the same chart or table, or a group of mechanical comprehension items all referring to a single diagram. Another example is a test consisting of a number of matching exercises, each of which consists of a homogeneous group of items. A group of items such as any of these may well seem to share specific content (category II B). In that case it will be preferable to put all the items in a single group into one half-score, and base the two half-scores on alternate groups of items. This procedure reduces one of the objections to the odd-even items procedure. Insofar as there is any difference between the two, this procedure may be expected to give a somewhat more conservative and a more appropriate estimate of reliability than odd vs. even items.

First vs. second half as basis for splitting test

In order to avoid the possibility of correlated errors of measurement arising from relatively short-time fluctuations in performance, the procedure has sometimes been adopted of correlating score on the first half of a test with score on the second half. This introduces obvious difficulties whenever the test is not quite homogeneous in content or when items become harder as the test proceeds. If there is a systematic shift in content or function from the beginning to the end of the test, the first and last half are clearly not equivalent. In the ordinary test of aptitude or achievement, this procedure would seem likely to give a less satisfactory approximation to equivalence in the two half-scores than would the odd-even items procedure.

Even when the formal content of a test is homogeneous from beginning

to end, as in the case of a series of trials in some complex motor task, the function may change qualitatively for the individual subject. That is, with continued practice the individual may find that the demands and character of the task change. What was initially a problem in discovering correct responses and procedures may, with practice, have changed to a task in developing maximum speed and precision of motor control. Thus, even identity of the external definition and formal requirements of the task cannot guarantee equivalence of the task as faced by the subject. On the one hand, this raises some question as to the interpretation of first vs. second half procedures (or of retest procedures) for evaluating reliability. At the same time it makes these procedures very interesting for comparison with an odd-even split-half determination. The amount that the first vs. second half measure is lower than that from odd vs. even items provides some indication of the extent to which either (*a*) momentary fluctuations worked to inflate the odd-even coefficient or (*b*) a progressive change in function with practice worked to lower the first vs. second half values. No simple way of discriminating between these two effects appears available for the traditional single test period. When the data consist of a series of trials of a learning problem, introduction of a considerable time interval between trials would appear to minimize the first factor.[6]

4. Analysis of Variance among Items

Any procedure for subdividing a total test into a particular two halves must be chosen somewhat arbitrarily from among the very large number of possible ways of making that subdivision. With that in mind, several workers have developed procedures to make use of all the information about consistency of performance from item to item within the test and thus provide a unique estimate of internal consistency. The procedures and basic formulas which are discussed here were first presented by Kuder and Richardson (*14*). The derivation of the most generally useful formula (Kuder-Richardson formula #20) has subsequently been carried out on the basis of less restrictive assumptions by Jackson and Ferguson (*11*), and has been related directly to the approach through analysis of variance by Hoyt (*10*).

[6] A technique has been suggested by Cureton for estimating the reliability for a learning function or performance which shows change with practice at a particular moment in its life history. If a series of trials are given, the correlations may be obtained between each possible pair of trials. If all correlations separated by a fixed number of trials are separately averaged (that is, the average of adjacent trials, those separated by one intervening trial, those separated by two, and so forth), a function may be plotted showing the relationship of average correlation to trial separation. This will generally be found to drop as the separation between the trials increases. By extrapolating backward to zero separation, one gets an estimate of the reliability (on the average) for a single trial.

This general approach yields a type of reliability estimate analogous to those obtained from subdividing a test and has many of the same characteristics and limitations. *In particular,* these procedures are not applicable to a test which involves the element of speed and is administered with a single time limit. The assumption is implicit in the method that the individual has attempted each item. Item characteristics, such as item difficulty, item variance, and item intercorrelations, become quite meaningless when any appreciable percentage of the group has not had time to read and attempt the item. For example, if omits are treated as wrongs, item intercorrelations toward the end of a speeded test become grossly inflated by the common group of subjects who never attempted the items. If individuals having omits are not included in the population from which item intercorrelations are computed, there is no uniform population upon which statistical analyses can be based. In other words, the same difficulties with speeded tests which were met in the case of a split test are again encountered in those procedures which analyze consistency of performance on single items. Consistency of performance cannot be evaluated unless the subject had an opportunity to perform.

Again, analysis of the consistency between items or trials of a test provides an estimate of consistency at a specific time. The temporary factors which were grouped in categories III and IV A of Table 8 remain relatively constant for each individual during a single test period, and are, therefore, considered as systematic rather than error variance. No estimate of the day-to-day consistency of the individual is possible with those procedures.

The most generally useful of the formulas for estimating reliability from the relationship of total test variance to item variance is Kuder-Richardson formula #20. This formula is

$$r_{tt} = \frac{n}{n-1}\left\{\frac{s_t^2 - \sum_{i=1}^{n} p_i q_i}{s_t^2}\right\}, \tag{6}$$

where r_{tt} = reliability of the total test,
n = number of items in the test,
s_t^2 = variance of the total test,
p_i = proportion passing item i,
$q_i = 1 - p_i$.

Perhaps the most general derivation of this formula is that by Jackson and Ferguson (*11*). They define a pair of equivalent tests, T and T′, as tests for each of which the variance is equal, for each of which the average

item covariance is equal, and for which the average covariance of items within a test is equal to the average covariance of items of one test with items of the other. That is

$$s_t^2 = s_{t'}^2$$

and

$$\overline{r_{ij}s_is_j} = \overline{r_{i'j'}s_{i'}s_{j'}} = \overline{r_{ij'}s_is_{j'}}.$$

If this definition is accepted, formula (6) can be derived quite simply. We must examine this definition, therefore, to see what is implied by it, so that the uses and limitations of the formula may be clarified.

The first implication in the above definition is homogeneity of content. The average $\overline{r_{ij'}s_is_{j'}}$ includes terms of the type $r_{ii'}s_is_i$, that is, covariance of pairs of matched items from the two parallel tests. Within the single form of the test, of course, these matched pairs do not appear. If the average within-test covariance is to be as great as the average between-test covariance, it must be accepted that the average covariance of these pairs of matched items is no greater than the average covariance of other item pairs. In proportion as heterogeneity appears in the test from item to item or from one group of items to another, Kuder-Richardson formula #20 will provide an underestimate of the correlation between equivalent forms. Since the terms in the above average which are based on matched item pairs are only n out of a total of n^2 terms, and since the difference in average covariance between matched and unmatched terms will often not be very great, the underestimation can be expected to be rather small in magnitude.

A second implication of this derivation of Kuder-Richardson formula #20 is that the test be essentially a power test. If there is a speed factor such that individuals fail to answer an appreciable number of items, a common factor is generated among the items of a single form of the test such that the average covariance among items within a test is raised above the average covariance between items on different forms of the test. That is, there will then be a substantial pool of individuals who necessarily have failed both items of a pair because they have not had time to attempt either of the items. On parallel forms of a test, where the testing operations are experimentally independent, different persons may have failed to complete the items, and so the covariance between a pair of items may be less. Therefore, on a speeded test an estimate of consistency based on a single testing will be too high. No general statement can be made as to the magnitude of this effect, but in measures in which individual differences are almost entirely a matter of speed, the overestimation may be

substantial. This is the same spurious source of reliability which we have previously noted in the case of split-half reliability estimates.

A third implication, or perhaps limitation of the derivation of this formula is that changes in the individual from one time to another are not considered as a source of variation. If they are a significant source, then obviously the average covariance between items within a single test form, all of which are administered to the individual at the same time, will be higher than the average covariance between items on different forms given at different times. In proportion as this variance over time increases, the Kuder-Richardson formula #20 will overestimate the reliability obtained from separate testings. It provides an estimate of precision in appraising the performance of an individual at a particular point in time.

In summary, the factors which serve to distort the reliability estimates from Kuder-Richardson formula #20, and the nature of their effects, are as follows:

1. *Heterogeneity of item content* operates to lower the value obtained by the Kuder-Richardson formula, but probably produces rather a small distortion in most cases.

2. *The speed factor* operates to raise the value obtained with the Kuder-Richardson formula, whenever there are unattempted items, by an amount which is unknown but probably becomes quite substantial in highly speeded tests.

3. *Diurnal variation* which is not represented in the Kuder-Richardson estimate (or any other based on a single administration), has the result that the values obtained are likely to be high by an unknown amount as estimates of consistency over a period of time.

It would appear, therefore, that this formula is most serviceable in estimating the consistency of performance on a relatively homogeneous power test when interest is focused on consistency of performance at a particular point in time.

In addition to the original development by Kuder and Richardson (*14*) and one by Jackson and Ferguson (*11*), this same essential formula has also been derived by Guttman and by Hoyt. Guttman (*8*) derived this expression as one of a series of expressions to provide minimum estimates of reliability, but his estimates are minimum only if one considerably restricts the types of error variance in which one is interested.[7] We have

[7] Guttman has developed a series of expressions which purport to represent "lower bounds" for the reliability coefficient. One of these expressions has been represented in Table 9 and one is identical with Kuder-Richardson formula #20.

These expressions are based on a single test administration, and purport to be minimum estimates of what the correlation would be between two administrations of the test. In

seen above how under certain very important circumstances the Kuder-Richardson formula may become an overestimate.

Hoyt (*10*) attacked directly the problem of estimating test reliability from consistency of individual performance upon the items of a test by analysis of variance techniques. He assumed that the score of an individual on a test may be divided into four independent (mutually uncorrelated) components, as follows: (1) a component common to all individuals and to all items; (2) a component associated with the item; (3) a component associated with the individual; (4) an error component that is independent of 1, 2, and 3. It is assumed further that the error component of each item is normally distributed, that the variance of the error component is the same for each item, and that the error components for any two distinct items are uncorrelated. When these conditions are met, it is possible to analyze the variance in test scores into the variance contributed by each of the last three components. (The first component is a constant for all items and all individuals, and hence is not a source of variance.) Reliability may be estimated from the expression

$$\text{Reliability} = 1 - \frac{\text{Error variance}}{\text{Variance among individuals}}. \tag{7}$$

If data are available for a total of N students on each of n items, the situation may be illustrated as follows:

	Items			
Student	1	2	$3 \cdots n$	Scores
1				t_1
2				t_2
3				t_3
.				.
.				.
.				.
N				t_N
Totals	p_1	p_2	$p_3 \cdots p_n$	$\sum_{i=1}^{n} p_i = \sum_{i=1}^{N} t_i$

In the above table, the t's represent scores of individual students, while

interpreting Guttman's formulas, however, it is essential to note the manner in which he defines reliability. Guttman explicitly eliminates from the sources of error with which he is concerned both (*a*) diurnal variation and (*b*) variation due to the sampling of items. It is only for a type of reliability which refers to performance on a particular set of items at a particular moment in time that his formulas provide "lower bounds."

the p's represent numbers of correct responses on the particular items of the test. The sum of squares "among students" is

$$\frac{1}{n}\sum_{i=1}^{N} t_i^2 - \frac{\left(\sum_{i=1}^{N} t_i\right)^2}{nN}, \tag{8}$$

and the variance among students is this quantity divided by N minus 1. The sum of squares "among items" is

$$\frac{1}{N}\sum_{i=1}^{n} p_i^2 - \frac{\left(\sum_{i=1}^{n} p_i\right)^2}{nN}, \tag{9}$$

and the variance among items is this quantity divided by n minus 1. The total sum of squares is

$$\frac{\left(\sum_{i=1}^{N} t_i\right)\left(nN - \sum_{i=1}^{N} t_i\right)}{nN}, \tag{10}$$

that is, the number of correct responses times the number of incorrect responses divided by the total number of responses. The residual or error sum of squares is the total sum of squares minus that attributable to the two systematic factors of individual and item. The error variance is the error sum of squares divided by $(n - 1)$ $(N - 1)$.

It has been indicated that the result obtained by Hoyt's procedure is identical with that from Kuder-Richardson formula #20, so nothing new is added so far as analysis of items of a test is concerned. The analysis of variance approach, however, appears useful for obtaining reliability estimates from items or trials which are scored with a range of scores, and not merely as "passed" or "failed." Thus, where several trials of a psychomotor test had been given, it would be possible to analyze the variance in performance into a portion of variance associated with trials, a portion associated with individuals, and a residual or error variance.[8] A quite general formula, applicable wherever a number of observations have been made upon each individual or specimen has recently been presented by Horst (*9*). This formula is applicable to the case when the number of observations differs from one specimen to another, and so is quite general in character. Horst shows that it reduces to the Spearman-Brown or to the Kuder-Richardson formulas under specific conditions. Horst's formula is

[8] This type of analysis has been elaborated further by Alexander (*2*) with a consideration of problems of identifying and eliminating trend effects.

$$r = 1 - \frac{\dfrac{\sum \dfrac{\sigma_i^2}{n_i - 1}}{N}}{\sigma_M^2}, \tag{11}$$

where N = the number of persons,
n_i = the number of measures for person i,
σ_i = the standard deviation of these measures for person i, and
σ_M = the standard deviation of the means for N persons.

There may be some occasions upon which investigators will wish to use Kuder-Richardson formula #21. This is a simplification of formula #20, arrived at by assuming that all the items are of the same difficulty. The formula then becomes

$$r_{tt} = \frac{n}{n-1} \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2}. \tag{12}$$

In this formula $\bar{p} = \frac{M_t}{n}$ and $\bar{q} = 1 - \bar{p}$. Thus, the only values which are required in this formula are the number of items, the mean, and the standard deviation for the total test. When the item difficulties are not actually equal, the value yielded by this formula will be lower, and sometimes substantially lower, than that resulting from K-R #20. However, in a number cases of fairly long power tests, the differences between the two formulas have been found (*20*) not to be greater than .05.

In the Kuder-Richardson formulas, it has been assumed that total score on the test is simply the unweighted sum of the number of correct responses. If items are weighted differentially, if wrong responses are weighted, and particularly if failures to respond are treated differently from wrong responses, the formulas become somewhat more complex. Dressel (*5*) has presented a generalized form of formula (6) above, suitable for use when weights are applied differentially to items and to rights, wrongs, and omits. Changing notation somewhat, we let

p_i = proportion of correct responses,
$q_i = 1 - p_i$,
p'_i = proportion of wrong responses,
$q'_i = 1 - p'_i$,
a_i = weight applied to correct response on item i,
b_i = weight applied to wrong response on item i.

Assuming that omitted items are not weighted, we get

$$r_{tt} = \frac{n}{n-1}\left[1 - \frac{\sum_{1}^{n} a_i^2 p_i q_i + \sum_{1}^{n} b_i^2 p'_i q'_i - 2\sum_{1}^{n} a_i b_i p_i p'_i}{\sigma_t^2}\right]. \quad (13)$$

This formula is equivalent to formula (6) in that case in which rights and wrongs are variably and differently weighted and omits are unweighted. However, these omits are those which the individual has presumably attempted, but elected to omit. This formula is no more acceptable than the others for tests which, because of their speed element, include many nonattempted items.

At this point it is perhaps appropriate to compare explicitly the measures of consistency which are obtained by correlating scores on equivalent forms of a test, by whatever experimental operations these have been obtained, with those based on analysis of the items in a single test. The difference in what is accomplished in these two approaches is sufficiently great so that some writers have objected to any use of the term "reliability" for the latter type, insisting that it be spoken of as "internal consistency." The basic difference is found in the relationship of *homogeneity* of the functions measured to the two types of indices of consistency of performance.

In the case of the Kuder-Richardson formulas and derivatives thereof, homogeneity of function measured is the basic assumption. Each item in the test is considered to measure the same factor or the same weighted combination of factors as every other. Variation in the factors measured from item to item results in a lower index of consistency, since this difference in factors measured lowers the correlation between items in just the same way that error variance does. A necessary condition for obtaining a high consistency index in this case is, therefore, substantial correlation of each item with every other, that is, a substantial factor common to all.

In an estimate of consistency obtained by correlating equivalent test forms, by contrast, it is theoretically possible to get perfect consistency from test to test even though no item within a single test form has *any* correlation with any other item in that test. Thus, if one test, X, is made up of items x_1, x_2, x_3, and a parallel form, Y, of the test is made up of items y_1, y_2, and y_3, and if x_1 correlates perfectly with y_1, x_2 with y_2, and x_3 with y_3, then the two tests will have a perfect correlation with one another even if the correlations between x_1, x_2, and x_3 are all zero. The illustration has intentionally been made extreme and unrealistic to dramatize the point that in terms of consistency of performance from one testing to

another, a very mixed and heterogeneous measure may yield a very stable score. This type of stability is not represented in the Kuder-Richardson internal-consistency analyses.

Factors Influencing the Reliability of a Test

The reliability coefficient which will be obtained from administration of a particular form of test to an experimental sample depends upon three types of factors. First of all we have the series of experimental and statistical operations which are used to define reliability. The various possibilities under these headings have been considered in the previous section. Factors of a second type concern the group to which the test is administered. These factors are, of course, extrinsic to the test itself and serve primarily to confuse the problem of reliability estimation. It will be necessary to consider these factors in some detail in the following paragraphs. A third type of factors are those which characterize the test itself and the method of its administration. These intrinsic factors will be discussed after factors relating to the group have been dealt with.

Reliability as a Function of Group Variability

The feature of a group which is most significant in determining the reliability coefficient (but not the standard error of measurement) is the range of ability represented in the group. It will be remembered that total variance in test score can be divided into error variance and true variance, and that the reliability coefficient may be expressed in the form

$$r = 1 - \frac{s_e^2}{s^2},$$

where s_e^2 represents the error variance and s^2 the total variance. The error variance is that variance which would arise from repeated measurements of the same individual. (For the moment it is assumed that a single value of the error variance characterizes all individuals over quite a range in ability level.) The error variance is not concerned with and does not reflect the amount of variance *between* individuals, as the total variance does. Therefore, the numerator of the expression $\frac{s_e^2}{s^2}$ may be thought of as essentially constant in groups with different ranges of ability, while the denominator increases as the variability of the group increases. Assuming that the standard error of measurement is the same in groups of different ranges of ability,[9] the following equation may be used to express the rela-

[9] This assumption is often rather questionable, and so results obtained from applying formula (14) must be viewed as tentative. See the following section for a discussion of this point.

tionship between reliability coefficients in groups of different ranges of ability.

$$\frac{s}{S} = \sqrt{\frac{1 - R_{11}}{1 - r_{11}}}, \tag{14}$$

where s is the standard deviation in the less variable group,
r_{11} is the reliability coefficient in the less variable group,
S is the standard deviation in the more variable group,
R_{11} is the reliability coefficient in the more variable group.

A chart to facilitate application of this formula has been prepared by Rulon (*21*) and may be useful in certain cases where a research worker has many reliabilities to correct. The above formula permits the estimation of the correlation between the two forms of a test for a group other than that for which it was originally computed, if one form has been given to the new group and its standard deviation in the new group is known.

Sometimes it may be necessary to estimate the reliability in a basic population when one knows the reliability coefficient and standard deviation only for a sample which has been curtailed on some other variable. Thus, data on an achievement test might be available only for a restricted group which had been screened on a preliminary aptitude measure. One might wish to estimate what the reliability would have been in the total group had no screening been carried out. In this case the only data available on the *total* group are scores on the screening test. The basic formulas for this general problem were developed by Pearson (*17*) in 1903. A simplification, applicable to the reliability coefficient when the equivalent forms correlated have the same standard deviation and the same correlations with other variables, has been reported by Davis (*4*). The formula becomes

$$R_{11} = \frac{r_{11} + r_{12}\left(\dfrac{S_2^2}{s_2^2} - 1\right)}{1 + r_{12}\left(\dfrac{S_2^2}{s_2^2} - 1\right)}, \tag{15}$$

where r_{11} is the reliability coefficient for the curtailed group,
R_{11} is the reliability coefficient for the uncurtailed group,
r_{12} is the correlation between the curtailing variable, 2, and the variable under study in the curtailed group,
S_2 is the standard deviation of variable 2 in the uncurtailed group, and
s_2 is the standard deviation of variable 2 in the curtailed group.

Reliability as a Function of Average-Ability Level

In the immediately preceding discussion it was assumed that the error variance was the same at different ability levels. This assumption means equal variability in each row or in each column of the bivariate frequency distribution for two forms of the test. The assumption will not necessarily hold in particular cases, especially where the range of abilities under consideration is great.

Any particular test item functions to differentiate among individuals only over a limited range of abilities. For individuals below the minimum, the item will not function differentially because all will fail it. For those above the maximum, it will not function differentially because all will pass it. Between the minimum and maximum, an item is most effective in differentiating at the level where about half the individuals can pass it. (For fuller discussion of this point see chapter 9 on item analysis.) A test is a composite of items. How precise and consistent differentiations it will make at any score level depends upon the number of items which are functioning at that score level and how effectively they are functioning. The items in the usual well-made test are selected to give good differentiation over some particular range of scores, and may be expected to discriminate less consistently both below and above that range.

There are still other factors which may result in lower precision at extreme score levels. On the one hand, the nature and instructions of a test may be such as to cause persons of low ability to rely very heavily upon guessing to determine their answers. When this is the case, error variance will be greater at the lower score levels. In another type of situation the amount done by persons of low ability may be very small compared to the amount for persons of high ability, and as a result the variability may be small because of the restricted range of total scores. Consider a test of speed of reading applied to one group of slow readers able to read an average of 100 words in the unit of time and to another group of fast readers able to read 1,000 words in the same unit of time. The variation from test to test in number of words read would seem almost necessarily to be less for the slow readers than that for the fast readers, and the error variance would therefore be less. Whether the reliability coefficient would be higher for the slow readers would, of course, depend upon the total variance of scores in the two groups.

Enough has been said to indicate that either the standard error of measurement or the reliability coefficient may vary with the average-ability level of the group being studied. The variation with ability level is not statable in general terms and will need to be determined

empirically for any given measuring instrument. It depends on the structure of the particular test in terms of item difficulties and item intercorrelations. The possibility of such a variation should make the test user very hesitant to apply data on reliability obtained from a group at one score level to a group at a radically different score level. In cases in which the range of ability covered is rather large and in which the sample is of sufficient size to warrant more detailed analysis, it may be appropriate to determine the standard error of measurement at several different score levels or at critical score levels.

A convenient way of estimating the standard error of measurement at a particular score level is to take all the cases who have the specified total score. If the papers for these individuals are separately scored for odd and even items, or some other split into half-tests, an estimate of the standard error of measurement is provided by the standard error of the differences between the scores on the two halves of the test. This value is obtained by finding the actual difference between score on the odd and the even items for each individual and computing the standard error of this distribution of differences. It can also be computed from scores on a pair of equivalent tests. In this case one takes all cases whose *average* score on the two forms falls within specified narrow limits. For each individual, the difference between score on the two forms is determined. The standard error of this distribution of scores is computed, and must be divided by $\sqrt{2}$ to give the required standard error of measurement. It has been shown for various tests that this standard error of measurement varies appreciably from one score level to another, and that the assumptions and, therefore, the applicability of formulas (14) and (15) are not warranted.

Intrinsic Factors Affecting Reliability

To the test constructor the important question is: What characteristics of the test itself make for reliability? In responding to this question, we must consider (*a*) the characteristics of the single items, trials, or measurements of which the total score is composed, (*b*) the number of items, that is, the length of the test, and (*c*) the conditions under which the items are administered.

The typical test in education is built up of separately scored items, and the typical score is a linear combination of the separate item scores. That being the case, the correlation between groups of items can be expressed in terms of the correlations of each component item with itself and with the other items. Following the standard formula for the correlation of sums, the correlation between one score based on a group of *a* items and

another score based upon a different group of B items is approximately given by the expression[10]

$$r_{(1+2+\cdots+a)(\mathrm{I}+\mathrm{II}+\cdots+B)} = \frac{\sum_{1}^{aB} r_{pQ}}{\sqrt{a + \sum_{1}^{a^2-a} r_{pq}}\ \sqrt{B + \sum_{1}^{B^2-B} r_{PQ}}}. \qquad (16)$$

That is, the numerator is the sum of all the terms in the matrix aB' of correlations between items in the two parts, while the denominator is the square root of the product of the two matrices aa' and BB' of self-and-cross-correlations of items within the same part. (The self-correlations are, for this purpose, taken as unity.) When the a items represent one form (or half) of a test and the B items represent another equivalent form (or half) of the test, this formula is a way of expressing the reliability coefficient for the test (or half-test).

From inspection of this formula it can be seen that in the limiting case where all the item intercorrelations are zero the reliability of the test will also be zero, and at the other limit where the item correlations are all unity the reliability will be unity. In general, the higher the item intercorrelations, the higher the reliability of the test. However, the relationship is a complex one. If paired items or groups of items in the two forms of a test have high correlations, it is possible to obtain a very high reliability coefficient even though the correlations of different items within each test and of non-paired items in the two test forms are quite low. Thus, two forms of a test might be constructed, each containing one half vocabulary items and one half block-counting items. Even if the correlation between vocabulary and block-counting items were zero, the reliability of the test might still be high, theoretically even 1.00, provided the vocabulary items in one form had sufficiently high correlation with the vocabulary items in the other and the block-counting items in one had sufficiently high correlation with the block-counting items in the other. We may conclude, therefore, that a high general level of item correlations is a guarantee of reliability in the test, but that when reliability is defined in terms of equivalent forms it is still possible for reliability to be high in spite of low item intercorrelations.

The size of item intercorrelations depends in part on item reliability and in part on homogeneity of function measured from item to item. Item reliability is, perhaps, more a theoretical than a practical concept. It is

[10] This expression would be exact if all items were of the same difficulty, or if some other procedure were followed to give all items the same standard deviation.

conceived of in the same way that test reliability is, that is, as the ratio of true-score variance to total variance, but there are difficulties in establishing a set of empirical procedures for estimating it. Retesting with the same item after an interval and determining the correlation between performance on the two testings is a possibility, but the factor of memory of previous response makes this solution of the problem rather unsatisfactory. The other possibility is to prepare an "equivalent form" of the item, but it is usually much less clear than item A' is truly equivalent to item A than that test X', made up of a large pool of items varying in difficulty and details of content, is equivalent to test X, made up of a similar pool of items. Because of the unsatisfactory nature of both these sets of operations for defining item reliability, workers in test development have tended to avoid any statistics which require an evaluation of the reliability of single items.

A group of items would be considered homogeneous in function measured if they all had in common the same non-error variance. The proportion of common variance, in this sense, is best shown by the tetrachoric correlation between two items. If this correlation reaches the limit set by the reliabilties of the separate items, we may conclude that the items in the test are perfectly homogeneous in content. In practice, of course, it is usually impossible for us to determine how closely the item intercorrelations approach this limit, because we do not have any satisfactory estimates of the reliabilities of the separate items. In theory, item intercorrelations may fail to be high either because the individual's behavior on a single item reflects chance factors, or because the items are heterogeneous in the functions which they measure. In practice, this distinction often breaks down, since all that is known is the correlation between items, or more often the correlation between an item and a pool of other items. The correlation of item with test is analogous to the first centroid factor loading of a variable in a factor analysis (cf. Richardson [*19*]). The part of the variance which is not accounted for by this factor may be either error variance (item unreliability) or variance in other factors (item heterogeneity). In these cases the average level of item intercorrelation is an expression of the degree of consistency of performance throughout the test, this consistency reflecting both the consistency of behavior in response to a single item and the homogeneity of function from item to item. The dependence of test reliability upon interitem correlations is seen most clearly in the original development of the Kuder-Richardson formulas (*14*), which were developed from the formula for the correlation of sums, making the assumption of complete homogeneity of function from item to item.

High item reliability would appear to be unequivocally desirable. That is, given that exactly the same functions are measured by two items, the one which measures those functions with greater consistency and precision (less error variance) is to be preferred. A high level of homogeneity of items may, however, be a mixed blessing. While greater homogeneity will generally result in greater reliability, it may do so at the cost of validity. Just as low correlations among tests permit higher multiple correlation from a combination of those tests, so low correlation among items permits a higher validity for the total test composed of those items. Where a single test is used for purposes of practical prediction, a great deal of concern for homogeneity of test materials in order to achieve high reliability is probably a mistake.

There are other testing purposes for which obtaining homogeneous test material will be a matter of more concern. In analytical studies of factors underlying individual differences in human behavior, concern for homogeneity in the items of each testing instrument seems appropriate. Again, whenever a battery of tests is being used for personnel classification, that is, prediction of a number of criteria, concern for the simplicity and homogeneity of the items in the separate tests in the battery is a legitimate technique for holding down the correlations among the different tests, and thus permitting more differential classification. Similarly, in the clinical use of a series of tests, the score on a test made up of homogeneous items, and ones which are factorially relatively simple, can be interpreted with more confidence than can a score which is a varied composite of a number of functions.

A slightly different concept of homogeneity has been expressed by Loevinger (*15*), and offered as a substitute concept for reliability in the development and appraisal of tests. A perfectly homogeneous test is defined as a test in which, after the items have been arranged in order of difficulty, an individual will pass all the items up to a particular item A and fail every item beyond that one. This conception of homogeneity incorporates both the concept of homogeneity of function and the concept of item reliability from the preceding paragraphs. That is, to meet this criterion, every item would not only have to measure the same function as every other but also elicit perfectly consistent behavior from the individual. It implies not only an absence of specific factor variance, but also an absence of error variance. This usage of homogeneity seems, therefore, to be a somewhat confused and perhaps unfortunate one, particularly so since we may wish to take a rather different attitude toward consistency of behavior and complete homogeneity of function measured.

Loevinger (*15*) has developed an index to express the type of homogeneity which is represented by her definition as previously stated. The formula is

$$H_t = \frac{V_x - V_{het}}{V_{hom} - V_{het}}, \tag{17}$$

where H_t = the homogeneity index for the test,
V_x = the obtained variance of the test,
V_{het} = the variance of a perfectly heterogeneous test with that distribution of item difficulties, and
V_{hom} = the variance of a perfectly homogeneous test with that distribution of item difficulties.

For estimating H_t, Loevinger develops the formula

$$\text{Est } H_t = \frac{N(\sum X_k^2 - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}{2N(\sum iN_i - \sum X_k) + \sum N_i^2 - (\sum X_k)^2}, \tag{18}$$

where N = number of persons tested,
X_k = score (number right) of person k,
N_i = number of persons passing item i,
i = order of difficulty of item i; $j > i$ if item j is more difficult than item i.

This formula yields an index which has a value of zero when a test is completely heterogeneous (that is, success on item i is entirely unrelated to success on other items), and a value of unity when a test is completely homogeneous (that is, success on item i is always accompanied by success on easier items). Other values fall in between these limits. Loevinger proposes that this index or one like it be used in place of the conventional reliability coefficient. However, this index is quite different from the usual reliability index in its properties. It gives very different numerical values. Gage and Damrin (*6*) report, for example, values of H_t of approximately .30 for a subjectively very homogeneous verbal intelligence test with a Kuder-Richardson reliability above .90. Furthermore, H_t showed no systematic increase with increased length of test, as, of course, the reliability coefficient did. If one accepted the maximizing of this homogeneity index as an objective of test construction, the distribution of item difficulties in a test would be very different from that yielded by current practice. A maximum homogeneity index would result from a rectangular or U-shaped distribution of item difficulties, rather than the bell-shaped item difficulty distribution which often seems to provide maximum reliability. All in all,

it seems doubtful that the index of homogeneity presents a useful alternative to the standard reliability estimates.

Test reliability is also a function of item difficulty. At any particular ability level, it is found that maximum reliability results when the items are of 50 percent difficulty level. Since in practice a test is almost always built with the purpose of discriminating over a range of levels of ability, and since the general estimate of reliability is an averaging of reliabilities at different score levels, maximum reliability is usually attained when items are spread in difficulty level over the range within which discrimination is to be made. Further discussion of item characteristics in relation to test reliability will be found in chapter 9.

A second major factor in the reliability of a test is the number of items. This is expressed in the generalized Spearman-Brown formula for estimating the reliability of a test of any length n from a test of unit length (formula [5] given on page 581). However, this formula assumes that the added items are equal in quality to those of the original test. That is, each added unit of length must show the same variability and same correlations with other units as does the original unit test. In addition to the problem of providing testing time and the question of stability of the function in the subjects tested throughout a long period of testing, the difficulty of producing further large numbers of equally good items must be considered. It has been shown that the reliability of many tests could actually have been increased by omitting a number of the items in the test. These are usually the items with the lowest item intercorrelations, since these are likely to be the most unreliable items. Lengthening a test is not a guaranteed way of increasing its reliability, if the additional material has lower item intercorrelations or lower item reliability than the original. The quality of the items, as discussed in previous paragraphs, may outweigh the quantity, and for some purposes a brief test built of the most consistent items may be preferable to a longer test of less carefully chosen materials.

Length of a test may be thought of not only as the number of items, but also as the number of options per item. Remmers and his co-workers (for example, *1*) have shown that the reliability of a test increases approximately as would be predicted by the Spearman-Brown formula when the number of response options is increased, provided that the added options are as attractive as those previously in the item. In a sense, a multiple-choice item acts as if it were the sum of $n - 1$ true-false items, where n is the number of response options. This relationship has been found with fairly small numbers of response options (that is, 2 to 5), but whether further increments in reliability would be obtained by further comparable

increases in the number of response options seems more questionable. This is, of course, in part a reflection of the difficulty of finding additional equally attractive response options. It is probable, also, that this relationship differs for different kinds of tests.

Finally, reliability is a function of the uniformity of testing conditions. Any variation in testing conditions from one test administration to another may be expected to be a source of variance in test performance. This variance must be considered error variance and will have the effect of reducing the reliability of the test. Several factors may be mentioned as contributing to uniformity of testing conditions.

One factor making for uniformity is the provision of adequate instructions and practice exercises for the test. Especially when the test performance is a somewhat novel one, subjects may be expected to spend an appreciable time becoming adapted to the test and perfecting their skills of dealing with the test materials. During this period of adaptation the individual's level of performance on the test is progressively changing, and this change may be taking place at a different rate for different individuals. During this period the function being measured is an unstable one in each individual, and changing levels of individual performance may be expected to lower reliability. It is desirable that as much as possible of this adaptation take place within the period of instructions and pretest practice exercise so that the performance during the period of the test proper may be as stable as possible.

Where time is a critical element in the test, accuracy of timing is, of course, an element in maintaining reliability. Differences from person to person or from subgroup to subgroup in the amount of time which they have been allowed for taking the test will be a source of variance in test score. Since these variations would ordinarily not be duplicated in another testing, they are to be thought of and treated as error variance. Similarly, where the test uses apparatus or equipment of any type, accurate maintenance and calibration of the equipment is a prerequisite of reliable measurement. Probably the largest-scale use of apparatus tests in a mass testing program took place in the Aircrew Testing Program of the Army Air Forces in World War II. This testing demonstrated that scrupulous care in the calibration and routine maintenance of apparatus was necessary if consistent testing conditions were to be maintained for all individuals. Even with the best efforts of testing personnel, apparatus differences and apparatus variation were still a matter of serious concern, and in some cases separate conversion tables were necessary for each "copy" of an apparatus test.

In certain cases the precise definition of the measurement to be obtained and the conditions of obtaining it represent an important aspect of uniformity of testing. In taking bodily measurements, for example, precise and unequivocal definition of the exact points on the body between which the measurements are to be taken constitutes an important condition of obtaining consistent and reliable measurements. In such a measurement as basal metabolism, the definition of those conditions which constitute the basal state for the individual represents a very large part of the procedure for measurement. In many cases then, particularly for measures other than paper-and-pencil tests, a complete and uniform statement of the operations and conditions for the measurement is a fundamental aspect of obtaining reliability.

Finally, mention may be made of standardization of motivation. Particularly in tests requiring consistent effort by or cooperation of the subject, standardization of motivation is very important. This becomes perhaps most acutely the case in measures of personality in which falsification of the results is quite easy. Developing definite and uniform motivation becomes one major function of the instructional and practice period.

It should be pointed out that many of the factors which have just been discussed will operate to lower test-retest reliabilities but to raise split-test coefficients. The split-test procedures will tend to allocate variance resulting from changes in testing conditions to systematic rather than error variance. This deficiency of split-test procedures has been noted earlier.

Interpretation of Estimates of Reliability

Relative vs. Absolute Measures of Precision

It was indicated at the beginning of this chapter that reliability or consistency of performance can be expressed either in absolute or in relative terms. Absolute consistency takes the form of the standard error of measurement. This is the standard error of the distribution of scores all of which are estimates of the same true score. It is given by the formula

$$S_e = S\sqrt{1 - r_{11}}. \tag{19}$$

This is essentially a rearrangement of formula (2) on page 566. The standard error of measurement is of particular value when we are interested in applying the information with regard to consistency to different groups. It has been shown in a previous section that the reliability coefficient depends upon the range of ability in the group from which the coefficient was determined. This makes it impossible to apply the coefficient directly to another group differing in variability on the trait in question or to compare

directly results from such different groups. The standard error of measurement is usually relatively independent of exact spread of scores, though it may differ in different parts of the range. It is reasonable, therefore, to expect the standard error of measurement to remain uniform in groups of approximately the same level of ability. This means that it is possible to apply that value directly to new groups which may differ considerably in variability from the group on which the standard error of measurement was originally determined. Where it is desired to apply reliability data to various different groups, the standard error of measurement has definite advantages.

A second situation which arises in which the standard error of measurement has unique advantages is when a test has been constructed with the specific aim of discriminating at a particular point or within narrow limits of the total range of ability. Thus, a test might have as its purpose the distinguishing of the top 10 percent of a group from the lower 90 percent, and the items might have been selected with the special purpose of differentiating at this point rather than over the whole range. In that case the standard error of measurement expressed in scaled scores should not be expected to be uniform over the whole range, but would be a minimum near the critical point. (It should be noted that if the increased accuracy in measurement at the critical point is secured by adding more items at the corresponding level of difficulty, the standard error of measurement expressed in *raw-score* terms will increase, although its value in scaled score units decreases.) The relevant estimate of that test's precision for the particular purpose for which it was designed would appear to be the standard error of measurement (in scaled scores) of individuals falling at or near the critical point. General correlational measures of performance over the whole range will provide only an average estimate which will fail to take acount of the specific purpose of the test.

Most of our discussion of reliability in this chapter has been concerned with relative precision of measurement and has been expressed in terms of the reliability coefficient. Certain limitations in the application of such a reliability coefficient have just been discussed. In certain situations, however, the reliability coefficient is to be preferred to an absolute measure of consistency such as a standard error of measurement. This is true whenever it is necessary to make comparisons between different tests for the same sample of cases. In general, no two tests have raw scores which are expressed in comparable units of measurement. This makes any direct comparison of abolute measures of consistency from test to test impossible. There is no guarantee that a standard error of 5 points on one test has meaning which is the same as or even similar to a standard error of 5 points

on a different test. In such a situation the only comparison of reliabilities which has meaning is the comparison of measures of relative precision. Consistency of placement within the group, if the group is the same for both tests, provides a meaningful basis for comparing the precision of measurement in two or more different measuring instruments.

It will be true in general that absolute measures of precision will have meaning only for the person who has a good deal of familiarity with the units in which the measurement has been carried out and of "feel" for the range of scores to be expected in such a measure. Thus, the statement that "the standard error of measurement of Stanford-Binet mental ages for ten-year-olds was found to be 6 months" will have meaning to the reader only insofar as he has some concrete sense of how much difference 6 months of mental age amounts to either in the type of performances represented or in relation to the variability among ten-year-olds in general.

Information Necessary for Interpretation of Reliability Estimates

In general, for reliability data to be meaningful a good deal more information must be supplied to the reader than a mere reliability coefficient or standard error of measurement. For maximum usefulness the research worker and test developer should report at least the following items of information in presenting data with regard to reliability estimation.

a) *The particular set of operations by which consistency of performance was estimated.* This should specify whether the estimate was based upon administration of equivalent test forms, repetition of the same test, subdivision into part scores of a single test administered as a unit, or analysis of the consistency of response to individual test items. If the reliability estimate is based upon two separately timed test administrations, the report should indicate how much time elapsed between them. If the estimate is based upon subdivision of a single test, the report should specify on what basis the two half-tests were assembled.

b) *Description of the group tested.* In reporting reliability data, any available characterization of the group used in the experimental testing should be presented. This description should include such things as grade or range of grades, age range, sex, general background, and economic characteristics of the sample. Any descriptive material which will help the reader to obtain a picture of the group and to estimate its comparability with groups with which he is concerned should be provided.

c) *Statistical characteristics of the group.* Any report of reliability data should present not only background descriptive characteristics of the group studied but also the basic statistical constants of the group for the test in question. Of particular importance here are the number of cases, the mean score, and especially the standard deviation of scores on the test.

d) *Adequacy of sampling.* It is important to report data indicating the adequacy of the sampling procedures employed. A reliability coefficient reported for a test is always an estimate, based on a sample, of the reliability of the test for a given population. It is, therefore, very important to know what kind of sample (random, stratified, haphazard, etc.) was employed, and how large is the estimated sampling error in the coefficient obtained. It is particularly important, in sampling from school populations, to recognize that the school rather than the pupil is the true unit of sampling. (See pages 253–54.)

The Value of Reliability Data

We must now consider the practical significance of data with regard to reliability. It is important to understand both what reliability tells us about a test or other measuring device and especially what it does not tell us. Some statement with regard to acceptable levels of reliability should perhaps be given as a guide to the use of tests.

In general, the important thing for the user of tests to remember with regard to reliability data is that while these data provide an estimate of the precision with which the test measures, they provide no information at all as to whether it measures what the test user wants to measure. Data concerning what the test measures must be sought from some source beyond the statistical analyses of the test itself. This external source may be the statistical comparison of test scores with some other outside criterion. It may be the critical, rational analysis by the research worker of the content of the test and of the nature of the function which it is desired to measure. For full discussion of these problems, see chapter 16 on "Validity." Obviously the determination that one is in fact measuring those functions which one desires to measure is basic to any subsequent inquiry into the precision of measurement. Data on reliability, therefore, supplement an initial judgment with respect to the validity of a test and provide information on the precision of measurement once the content of measurement has already been determined.

In the evaluation of achievement test scores, reliability data may often appear to be the most basic statistical data with regard to the test. This is

the case because in tests of this sort validity tends to be established by other than statistical means. Validity becomes a question of appropriateness of content to the course subject matter and appropriateness of form of test to the course objectives. These points are often evaluated by rational rather than empirical methods. Once the test constructor has arrived at the outline of specifications for his test in terms of the rational type of approach which we have discussed, his interest shifts to precision of measurement and to statistical estimates of reliability.

In the case of aptitude tests developed for predicting some specific criterion of academic or job achievement, statistical estimates of reliability tend to become quite secondary to statistical evidence of validity. The reliability of a test usually becomes of interest in this context when one raises the question of the fruitfulness of trying to increase test validity by lengthening the test and thereby obtaining higher reliability. Knowing the reliability and validity of a test of unit length, it is possible, under the assumptions of formula (5), to estimate the validity which would be obtained for a test of length n. The formula is

$$r_{0(n)} = \frac{r_{01}}{\sqrt{\frac{1}{n} + \left(1 - \frac{1}{n}\right) r_{11}}} = \frac{r_{01}}{\sqrt{r_{11} - \frac{r_{11}}{n} + \frac{1}{n}}}. \qquad (20)$$

Inspection of this formula shows that the validity increases only as a function of the reciprocal of the square root of the reliability, so that it is generally true that increments of validity from lengthening an already moderately reliable test will not be great. Where an aptitude test has been developed for possible use with a number of jobs, appraisal of reliability does become a significant preliminary analysis in advance of validation for the several job categories.

Reliability data are also of some significance in evaluating criterion measures. A crucial point here is that the criterion must have a minimum reliability different from zero if any prediction of that criterion is to be possible. Though reliable criterion measures are eminently desirable, since they reduce the chance errors of measurement and consequently make for greater accuracy in the determination of relationships between predictor and criterion, high reliability in a criterion is less important than maximum relevance of the criterion to the goals in which one is ultimately interested.

Information about the reliability of the criterion is also important in enabling us to estimate the correlation which would be obtained between a predictor variable and a hypothetical perfectly reliable criterion.

This is the correlation in terms of which the obtained estimate of test validity should be appraised.

Interpretation of Reliability Coefficients of Different Sizes

One may be led to inquire how reliable a test must be to be used for different types of testing projects. One answer to this question was offered by Kelley (*12*) and has since been widely quoted. Making the assumption that for a test to be useful it must permit discriminations of a difference as small as 0.26 times the standard deviation of a grade group with chances five to one of being correct, Kelley arrives at the following as the minimum correlation for several purposes.

a) To evaluate level of group accomplishment .... .50
b) To evaluate differences in level of group accomplishment in two or more performances .... .90
c) To evaluate level of individual accomplishment .... .94
d) To evaluate differences in level of individual accomplishment in two or more performances .... .98

It must be recognized, however, that these values are arbitrary, being derived from the above assumptions as to what it would be reasonable to expect a test to do in the way of discrimination between individuals and groups. How low a reliability one is willing to accept in any given case depends upon the practical values which are involved in that particular case. If some action must necessarily be taken and only unreliable measures are available as a basis for action, one may have to make the best of an unsatisfactory situation and use the most reliable of the available measures even if it has a reliability coefficient of only .40 or .50. Thus, if an industrial organization required an appraisal of leadership ability to use in selecting from among its workmen those most likely to be successful as foremen, in order to provide them specialized training, and if no procedures for estimating leadership were available which had reliability above .50, the company would then necessarily use a procedure with reliability of .50 or less. The practical situation often does not permit waiting until procedures have been developed which meet a statistician's ideal of reliability. In practice, the research worker must develop a sense of the range of values which have ordinarily been obtained in different types of measurement and feel satisfied when he surpasses the values which are typical of the field in which he is working and dissatisfied when he falls below them.

For interpreting coefficients of reliability, it is important to remember that correlation coefficients are not to be thought of as percents. A coefficient of .50 does not mean that the error in measurement has been cut in half. The reduction is much less than that, and formula (19) shows that the standard error of measurement in this case is 71 percent of the value

TABLE 10

RELATIONSHIP BETWEEN RELIABILITY COEFFICIENT AND STANDARD ERROR OF MEASUREMENT

Reliability Coefficient	Relative Size of Standard Error of Measurement
.00	1.000
.25	.865
.50	.707
.70	.548
.80	.447
.90	.316
.95	.224
.98	.141
.99	.100

for a correlation of .00. The relative size of the standard error of measurement for a number of sample values of the reliability coefficient is shown in Table 10. It is a healthy corrective to undue enthusiasm about reliability coefficients of .80 or .90 to look at this table and note that the standard error is reduced only to approximately 45 percent and 32 percent of the value for zero reliability. It is apparent from Table 10 that a substantial variation in scores for an individual is possible even with tests of high reliability.

Estimation of True Scores

Given two sets of equivalent scores, each set being an estimate of the same underlying characteristic of the individuals measured, it is possible to make various estimates concerning the *true score* of the individuals in question. The true score for an individual may be defined as the score which would be obtained from the average of unbiased independent measurements of that individual if the number of applications of the measuring device were increased without limit, that is,

$$X_\infty = \frac{1}{n}\sum_{1}^{n} X \quad \text{when} \quad n \rightarrow \infty$$

where X_∞ represents the true score on variable 1.[11] If two sets of measurements are unbiased and independent, in that the errors of measurement

[11] The subscript ∞ will be used to designate a true score in variable 1, and the subscript ω to designate true score in variable 2.

average zero and are uncorrelated for the two sets, then a number of statistics relating to true scores may be estimated from them.

The correlation between a single set of fallible measures and the set of true scores for the individuals in the group is the square root of the reliability coefficient[12] for that group, that is,

$$r_{1\infty} = \sqrt{r_{11}}\,. \tag{21}$$

This value has commonly been called the *index* of reliability, and the correlation between two fallible measures has been spoken of as the *coefficient* of reliability.

The variance of true scores in the group equals the reliability coefficient times the variance for the distribution of scores on the single test, and the standard deviation of true scores is the square root of the reliability coefficient times the standard deviation of scores on the single test. That is,

$$s_\infty^2 = r_{11}s_1^2 = r_{1\infty}^2 s_1^2 \tag{22}$$

$$s_\infty = s_1\sqrt{r_{11}} = r_{1\infty}s_1\,. \tag{22a}$$

The true score for an individual may be estimated by the regression of the true upon obtained scores. Expressing scores as deviations from the group mean, we have

$$\tilde{x}_\infty = r_{1\infty}\frac{s_\infty}{s_1}x_1 \tag{23}$$

$$\tilde{x}_\infty = r_{1\infty}\frac{r_{\infty}s_1}{s_1}x_1 = r_{1\infty}^2 x \tag{23a}$$

$$\tilde{x}_\infty = r_{11}x_1\,. \tag{23b}$$

That is, the best estimate of an individual's true score is that it deviates from the mean of his group by r_{11} times the deviation of his obtained score.

Estimates of true scores of individuals are usually of interest for research rather than administrative purposes. One use of such scores is in setting up matched groups for experimental purposes, whenever groups are to be chosen from parent populations with different amounts of the trait in question. Since the true scores of a sample selected from a population will always regress somewhat toward the mean of the population, and since samples from different populations will regress toward different population means, if true matching is to be obtained it must be on estimated true scores rather than on raw obtained scores. Estimated true scores are also of interest in studying relative variability of different samples, in setting

[12] Derivations of the fundamental relationships presented in this section may be found in Peters and Van Voorhis (*18*).

up conversion tables for translating scores on one test to another (see pages 750–60), and a variety of other scaling and technical problems.

We are sometimes interested in estimating the correlation between true scores on two or more different qualities from the correlations obtained between actual, and necessarily somewhat unreliable, measurements of those qualities. If an estimate of the reliability of the different measures can be obtained, in accordance with procedures discussed earlier in this chapter, then it is possible to estimate the correlation which would be obtained if it were possible to replace one or both of the unreliable measures by the corresponding true measures. If only one of the fallible variables is to be replaced by the true measure the formula becomes

$$r_{\infty 2} = \frac{r_{12}}{\sqrt{r_{11}}}. \tag{24}$$

If both variables are to be replaced, the formula becomes

$$r_{\infty \omega} = \frac{r_{12}}{\sqrt{r_{11} r_{22}}}. \tag{25}$$

This last is Spearman's familiar formula for correction for attenuation due to the unreliability of the tests.

The first of the above formulas might legitimately be used if it were desired to know how accurately a test would predict a perfectly reliable criterion measure of some quality. Thus, if the reliability of a rating scale for assessing certain personal qualities is found to be .36 and the correlation of some objective test with score on that rating scale is found to be .30, we have

$$\frac{.30}{\sqrt{.36}} = .50$$

In this case it is legitimate to think of .50 as being the correlation which represents the true effectiveness of the test for predicting the "line" criterion score represented by the fallible ratings. The relationship of test score to the perfect criterion may be thought more germane than the relationship to a rather unreliable criterion instrument. Comparisons of tests in different areas where criteria may be of quite different reliability may more legitimately be made on this basis.

It is also sometimes of interest to inquire what prediction of an existing criterion could be obtained from a perfectly reliable test. The correlation in this case indicates the theoretical ceiling which might be reached in test-criterion correlation if the test were lengthened without limit. Thus, a

test with reliability of .81 and correlation with criterion of .36 would, if made perfectly reliable, correlate with the criterion to the extent of .40, in accordance with the calculation

$$r_{1\omega} = \frac{.36}{\sqrt{.81}} = .40$$

This permits a quick estimate of the value of trying to increase the reliability of any test by merely extending it quantitatively. The illustrative example indicates that little gain in validity may be anticipated by increasing the reliability of an already moderately reliable test.

Correlations corrected for the unreliability of both measures should be reported circumspectly and interpreted with care. They are of interest in a number of types of theoretical analysis. They indicate what proportion of their non-chance variance two variables have in common and thus serve to clarify basic understanding of the structure of human abilities. They also indicate what is the maximum correlation which could possibly be obtained between a particular type of predictor and a particular type of criterion variable if the only change were to make both perfectly reliable. However, practical prediction must be done with existing fallible tests. To some extent it is misleading to present corrected correlations between hypothetical true measures. The prediction which we could achieve of a hypothetical, perfectly reliable criterion is apropos since it is that criterion which we are in fact trying to predict, but the prediction which could be achieved by a hypothetical, perfectly reliable test may be quite misleading because such a test is never available to us.

Special Problems in Reliability Determination

There are several special situations which raise particular problems in connection with the determination of reliability. These are situations which make it difficult to obtain two independent but equivalent measures. A consideration of some of these particular problems may be helpful in providing the reader with advance notice of their existence and some suggestions for dealing with them when they are encountered.

Reliability of Speed Tests

It has been noted in a previous section that certain sets of operations for determining reliability become inappropriate for a speeded test. In a speeded test, amount accomplished within the time limit is a fundamental aspect of performance. To obtain two independent estimates of the amount that the individual can perform within a limited period of time, two

separate time limits are necessary. No test procedure which is built upon a single administration with a single time limit can give two separate speed measures. This means that any meaningful reliability estimate for a speeded test must be based upon two separate test administrations, either of the same test or of equivalent forms. Of course, a test may be divided into two separately timed halves and the scores from these used. These separately timed halves become in effect two shorter equivalent tests.

Reliability of Difference Scores

Psychologists and educators often have occasion to use scores which represent the difference between two measurements. This is the case whenever one is dealing with gains or changes, as when one studies the change in IQ after a year in preschool or the change in reading comprehension from the beginning to the end of the ninth grade. It is also the case when one deals with the profiles or sections of profiles, and is concerned with differences between IQ and reading achievement, for example.

Considering two different measures, X and Y, and defining equivalent forms of these two tests as we did on page 575, it can readily be shown that the reliability of the difference score $X - Y$ is given by the expression

$$r_{x-y} = \frac{r_{xx} + r_{yy} - 2r_{xy}}{2 - 2r_{xy}} \tag{26}$$

where r_{xx} is the reliability of test X,
r_{yy} is the reliability of test Y,
r_{xy} is the correlation between X and Y.

Thus, if the reliability of a reading test is .90, the reliability of an intelligence test is .85, and the correlation between them is .75, the reliability of measures of difference between intelligence and reading ability becomes

$$\frac{.90 + .85 - 2(.75)}{2 - 2(.75)} = \frac{.25}{.50} = .50.$$

The reliability of the difference score is substantially lower than that of either of the two component scores. As the correlation between the two tests approaches the average of the two reliabilities, the reliability of the difference score approaches zero. The generally low, and somtimes vanishing, reliability of difference scores is something of which both research worker and clinician need to be acutely conscious. More than one investigator has wasted time in trying to determine the correlates of difference scores of near-zero reliability.

Reliability of Learning Functions

Determination of the reliability of a learning function presents certain particular difficulties. This is true because in the very nature of things learning changes the individual and it is impossible to repeat the same task. The same external task is no longer the same once the individual has learned. The preparation of equivalent forms of a learning task is a dubious enterprise at best. By what criteria shall the research worker evaluate the equivalence of two learning tasks? On the one hand tasks which appear superficially similar may face the subject with somewhat different requirements. On the other hand tasks which might have been thoroughly equivalent had it been possible to hold the individual without change may well have lost that equivalence once the individual has been exposed to the original learning. The more nearly alike the tasks are, the more likely it is that exposure to one will distort subsequent performance on the other.

In estimating reliability of a learning function, one is almost necessarily thrown back upon the procedure of subdividing the course of learning into segments and correlating some grouping of segments with some other. This procedure is, of course, analogous to the split-test procedures for determination of reliability in a standardized test. Many of the objections which arose with regard to this procedure in connection with tests are also appropriate here. In particular, one may raise the question as to whether the scores on successive trials represent independent estimates of the individual's learning performance or whether they do not also partake of common errors of measurement. That is, it is suggested that adjacent trials in a learning performance are influenced to a certain extent by the same chance combination of temporary conditions (variance in categories III and IV in Table 8). Insofar as this is true, the split-test procedures will tend here also to give exaggerated estimates of test reliability.

Reliability of Tests with Element of Insight or Discovery

There exists a certain range of intellectual tasks and functions in which success depends to some measure upon insight into the nature of the task and discovery of a technique of dealing with it. In these cases performance in the task has something of an all-or-none character. The individual who does not "get the idea" of the task tends to flounder ineffectually and to make very little progress and a very low score. When the insight occurs or the technique is developed, the performance changes abruptly to a very much higher level. Score on this type of a test is to a considerable measure a function of the particular point during the test period at which the in-

dividual got insight into the nature of the test task; insofar as the test shows this characteristic, any adequate estimate of reliability for it is very difficult indeed to obtain. Just as in the case of the learning function it is hardly possible to repeat the test or a parallel form of it. Once the individual has achieved an insight into the test, the same task no longer exists for him and repetition of the task is impossible. At the same time split-test procedures are unsatisfactory in the same way they are in a speeded test. This is due to the fact that score is dependent upon the time at which insight was developed and that length of time would affect both half-scores in the same way. In other words two half-scores are not independent measures. Preparation of a test which depends upon different but similar type of insight as an equivalent form of test might be urged. However, the guarantee of equivalence in this case seems rather unsure. No ready solution to the problems raised above seems now available.

Summary

Evaluation of Procedures for Reliability Estimation

It was pointed out in the opening section of this chapter that variance in performance on a measuring instrument arises from sources of many different types. At one extreme we have the persisting general qualities of the individual. At the other extreme we have the completely unpredictable hazards which we label "chance." In between we find a wide range of factors of varying degrees of generality and permanence.

A number of different sets of experimental and statistical operations have been devised for estimating the reliability or precision of a measuring instrument. It has been pointed out that these different sets of operations differ in the way in which they define the variance which is to be thought of as error. Types of variance which are effectively allocated by one set of operations to systematic variation in performance from one individual to another are allocated by other sets of operations to "error of measurement." At this point it seems well to make some evaluation of the appropriateness of the different possible procedures.

The set of operations which seems to be most generally defensible for estimating the reliability of a measuring instrument is the preparation and administration of two or more equivalent forms of a test. The objections to this procedure appear to be primarily practical ones of the labor involved rather than logical ones concerning the appropriateness of the set of operations. The question as to the interval between administration of the test forms must be answered in terms of the particular use to be made of

the resulting figures. For use in connection with a table of correlations among tests administered at a single testing period, immediate retest seems appropriate. For use in connection with predictions and evaluations extending over some period of time, the meaningful procedure would appear to be to retest with a similar time interval.

With certain types of test materials, repetition of the same test becomes an appropriate substitute for the use of equivalent forms. This is true whenever the sampling of content does not need to be considered as an appreciable source of variance in performance and when the materials are sufficiently homogeneous and nondescript so that specific memories will not carry over from one testing to the other.

Where economy of time makes it undesirable to prepare two full-length equivalent test forms, a very useful practical approximation to this may be achieved by preparing the original test in two equivalent, separately timed halves. This procedure adds only a limited amount of inconvenience to test construction and test administration, while at the same time defining reliability by essentially the same operations as using full-length equivalent test forms.

In general, procedures involving a subdivision of the items in a single testing instrument seem less satisfactory than those based upon equivalent forms. The procedures become entirely unsatisfactory particularly in any test in which speed is a significant element in score. Where the test is appreciably speeded, no meaningful estimate of reliability can be obtained from a single testing period with a single time limit.

If a test is to be split to yield two scores, the procedure logically most defensible would appear to be to split it into two equivalent halves, balanced in terms of difficulty and content. In many cases the procedure of allocating alternate items to the two halves of the test will be a rough-and-ready way of approximating this result. However, the approximation cannot be relied upon, particularly in the case of relatively brief tests.

In those cases in which there is reason to believe that a test is homogeneous in content, estimates of reliability derived from consistency of performance from item to item become logically equivalent to those resulting from split-test procedures. The Kuder-Richardson and Hoyt procedures can be substituted for split-test procedures in these cases with profit, since they make use of all the data about consistency of performance from item to item. When the material of the test is not homogeneous, these last procedures provide an underestimate of the correlation between equivalent tests, but indications are that in the case of K-R #20 this underestimation is usually small in amount. Any situation, however,

which invalidates the split-test procedure operates equally on these measures.

Reporting Reliability Data

As statistical indices of precision of measurement, measures expressed both in relative and absolute terms serve useful functions. The correlation coefficient is the usual index of relative precision. Despite the limitations stemming from its nonlinear character, its general familiarity in education probably makes it the most useful statistic of this type. Precision may also be expressed in score units by the standard error of measurement. In some cases, where groups tested are very large or where the test has been constructed especially with the purpose of discriminating at a particular critical point, it may be appropriate to determine the standard error of measurement at one or more particular levels.

It has been demonstrated that the value which is obtained for the reliability of a test is a function of a number of factors both in the operations by means of which it has been determined and in the group upon which the determination has been made. If data with regard to reliability are to permit of meaningful interpretation by the reader, it is the responsibility of the author to present completely the facts with respect to operations and group. A report of reliability data should cover at least the following points:

1. The specific set of experimental and statistical operations upon which the estimate of reliability was based.
2. A descriptive characterization of the group which is as complete as possible with regard to all elements which might affect the reliability coefficient.
3. The statistical characteristics of the group, particularly the number of cases, the mean, and the standard deviation on the measure in question, which might be expected to affect the resulting reliability estimate.
4. The adequacy of the sampling procedure employed.

Significance of Data on Precision of Measurement

It must be remembered at all times that reliability or precision in a measurement is not an end in itself but only a means to an end. The end in view is, after all, the measurement of something which is practically or theoretically important. Reliability in the measurement contributes to the practical, and in some cases the theoretical, value of the measurement. However, precision cannot add importance to the measurement of something which is fundamentally trivial. Educators stand in danger

of overvaluing reliability in their measurement procedures at the expense of real significance in what is measured. Within the scope of a significant problem or procedure for measurement, greater precision is something for which the worker should strive assiduously. However, it is rarely something for which he should make any appreciable sacrifice of fundamental validity in the definition of that which he sets out to measure.

Though reliability should never become a primary goal in the development of measurement procedures, data with regard to reliability should always be available as an aid in the interpretation of measurement results. Both in the theoretical and practical fields of measurement and evaluation, the critical interpretation of experimental results will almost always depend upon knowledge of the precision of the several tools used in the measurement enterprise.

Selected References

General Discussions of the Statistics of Reliability

Cureton, E. E. *Errors of Measurement and Correlation.* ("Archives of Psychology," No. 125.) R. S. Woodworth (ed.). Vol. 19, 1930–31. N.Y.: Columbia University.

Kelley, T. L. *Statistical Method.* New York: Macmillan Co., 1924.

Thurstone, L. L. *The Reliability and Validity of Tests.* Ann Arbor, Mich.: Edwards Bros., 1931.

Materials to Which Reference Is Made in the Text

1. Adkins, R. M., and Remmers, H. H. "Reliability of Multiple-Choice Measuring Instruments as a Function of the Spearman-Brown Prophecy Formula," *Journal of Educational Psychology,* 33: 385–90, 1942.
2. Alexander, H. W. "The Estimation of Reliability When Several Trials Are Available," *Psychometrika,* 12: 79–100, 1947.
3. Cronbach, L. "Test 'Reliability': Its Meaning and Determination," *Psychometrika,* 12: 1–16, 1947.
4. Davis, F. B. "A Note on Correcting Reliability Coefficients for Range," *Journal of Educational Psychology,* 35: 500–502, 1944.
5. Dressel, P. L. "Some Remarks on the Kuder-Richardson Reliability Coefficient," *Psychometrika,* 5: 305–10, 1940.
6. Gage, N. L., and Damrin, D. E. "An Experimental Study of Single-Trial Estimates of 'Reliability' and the Concept of Homogeneity." Unpublished MS.
7. Gulliksen, H. "The Relation of Item Difficulty and Interitem Correlation to Test Variance and Reliability," *Psychometrika,* 10: 79–91, 1945.
8. Guttman, Louis. "A Basis for Analysing Test-Retest Reliability," *Psychometrika,* 10: 255–82, 1945.
9. Horst, A. P. "A Generalized Expression for the Reliability of Measures," *Psychometrika,* 14: 21–32, 1949.
10. Hoyt, C. "Test Reliability Obtained by Analysis of Variance," *Psychometrika,* 6: 153–60, 1941.
11. Jackson, R. W. B., and Ferguson, G. A. *Studies on the Reliability of Tests.* ("University of Toronto Department of Education Research Bulletin," 1941, No. 12.) 132 pp.
12. Kelley, T. L. *Interpretation of Educational Measurements.* Yonkers, N.Y.: World Book Co., 1927.

13. ———. "Note on the Reliability of a Test," *Journal of Educational Psychology,* 15: 193–204, 1924.
14. KUDER, G. F., and RICHARDSON, M. W. "The Theory of Estimation of Test Reliability," *Psychometrika,* 2: 151–60, 1937.
15. LOEVINGER, J. *A Systematic Approach to the Construction and Evaluation of Tests of Ability.* ("Psychological Monographs," Vol. 64, No. 285.) Washington: American Psychological Assoc., 1947.
16. MOSIER, C. I. "A Short Cut in the Estimation of the Split-Halves Coefficient," *Educational and Psychological Measurement,* 1: 407–8, 1941.
17. PEARSON, K. "Mathematical Contributions to the Theory of Evolution XI. On the Influence of Natural Selection on the Variability and Correlation of Organs," *Philosophical Transactions of the Royal Society of London,* Series A, 200: 1–66, 1903.
18. PETERS, CHARLES C., and VAN VOORHIS, WALTER R. *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill Book Co., 1940. Chap. 7.
19. RICHARDSON, M. W. "Notes on the Rationale of Item Analysis," *Psychometrika,* 1: 69–76, 1936.
20. RICHARDSON, M. W., and KUDER, G. F. "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence," *Journal of Educational Psychology,* 30: 681–87, 1939.
21. RULON, P. J. "A Graph for Estimating Reliability in One Range Knowing It in Another," *Journal of Educational Psychology,* 21: 140–42, 1930.
22. ———. "A Simplified Procedure for Determining the Reliability of a Test by Split-halves," *Harvard Educational Review,* 9: 99–103, 1939.
23. SPEARMAN, C. "Coefficient of Correlation Calculated from Faulty Data," *British Journal of Psychology,* 3: 271–95, 1910.

16. Validity

By EDWARD E. CURETON
University of Tennessee

COLLABORATORS: Ruth E. Eckert, *University of Minnesota;* C. L. Shartle, Ohio State University; Phillip J. Rulon,[1] *Harvard University* .

THE ESSENTIAL QUESTION OF TEST VALIDITY IS HOW WELL A TEST DOES THE job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. Hence, we cannot label the validity of a test as "high" or "moderate" or "low" except for some particular purpose. The Minnesota Clerical Test, for example, is a highly valid indicator of proficiency in some few clerical jobs, it is moderately valid for many others, and its validity for still others is quite low. Other tests not including the word "clerical" in their titles are at least equally valid, on the whole, for most of the same purposes (group intelligence tests, for example), and combinations of tests of both types are almost always more valid for such purposes than are tests of either type taken alone. Moreover, the Minnesota Clerical Test has been demonstrated to be a fairly valid indicator of proficiency in certain types of factory inspection and mechanical work. Validity is always validity for a particular purpose. It indicates how well the test serves the purpose for which it is used.

Purpose, in turn, has at least two aspects. One of these concerns the function to be appraised; the other, the nature of the group in which the appraisal is to be made. If a vocabulary test is given to a group of eighth-grade children all of whom have had fairly equal and fairly considerable opportunities and incentives to learn the meanings of printed words throughout their educational careers, and all of whom come from homes which are more or less similar in their general cultural characteristics, the test is likely to be a reasonably valid indicator of verbal intelligence. If it is given to a group of eighth-grade children whose educational backgrounds are very dissimilar, it may be more valid as an indicator of the

[1] The major part of the section on "The Criterion," along with a considerable number of paragraphs in other sections of this chapter, was written by Dr. Rulon. The author, however, has rearranged the materials sent to him by Dr. Rulon, modified some of the passages, and added passages, changing some of the essential meanings of these materials. He must, therefore, assume sole responsibility for the viewpoints presented, but wishes to acknowledge the very great value of Dr. Rulon's contribution.

general quality of previous instruction in reading than as an indicator of verbal intelligence. If it is given to a fairly heterogeneous group of adults whose formal schooling is long past, it may be more valid as an indicator of reading interest than as an indicator of either verbal intelligence or quality of past instruction in reading. The validity of any test is its validity as an indicator of individual differences in some particular function among the members of some specified group. The validity of a test—or of any observational technique—for a given purpose may range from zero to almost perfect. The practical problems of validity are concerned with these matters of degree.

Definitions

Relevance and reliability

Validity has two aspects, which may be termed relevance and reliability. "Relevance" concerns the closeness of agreement between what the test measures and the function that it is used to measure. "Reliability" concerns the accuracy and consistency with which it measures whatever it does measure in the group with which it is used. To be valid—that is, to serve its purpose adequately—a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to measure. If we want to find out how well a person can perform a task, we can put him to work at that task, and observe how well he does it and the quality and quantity of the product he turns out. Whenever a test performance is anything other than a representative performance of the actual task, we must inquire further concerning the degree to which the test operations as performed upon the test materials in the test situation agree with the actual operations as performed upon the actual materials in the situation normal to the task. One way to do this is to make detailed logical and psychological analyses of both the test and the task. From such analyses we may be able to show that many or most of the test operations and materials are identical with or very much like many or most of those of the task, and that the test situation is intrinsically similar to that of the task. On the basis of this demonstration it might be reasonable to conclude that the test is sufficiently relevant to the task for the purpose at issue. Its reliability would then be investigated separately by some appropriate method chosen in accordance with the principles discussed in chapter 15.

Criterion measures

Logical and psychological analyses of the type suggested in the previous paragraphs are beset with many pitfalls, and the resulting estimates of

relevance are at best qualitative. A more direct method of investigation, which is always to be preferred wherever feasible, is to give the test to a representative sample of the group with whom it is to be used, observe and score performances of the actual task by the members of this sample, and see how well the test performances agree with the task performances. The task performances are termed "criterion performances," and the scores on these performances are called "criterion scores." They are the standard against which the usefulness of the test scores is judged.

Neither the test scores nor the criterion scores will be perfectly reliable. But if the criterion scores consist of completely unbiased judgments based on observations of perfectly representative samples of criterion performance, they will be perfectly relevant to the purpose at issue. Such perfectly relevant scores are termed "ultimate criterion scores," even though they may be only moderately reliable. The validity of a set of ultimate criterion scores is identical with its reliability, which depends in turn upon the number of observations on which each score is based.

The definition of validity

Though the validity of a test, as an indicator of the quality of performance on any given type of task, depends upon both its relevance and its reliability, the effectiveness of the test is not a function of the reliability of the criterion scores, which is largely a matter of the number of task performances of each member of the experimental sample which are observed and scored. Validity is therefore defined in terms of the correlation between the actual test scores and the "true" criterion scores. A "true" score is that part of any actual score which is not error of measurement. The quotation marks serve to emphasize the fact that any such correlation must be estimated rather than computed directly; the "true" score of a single individual cannot be obtained from any test or observational series of finite length. A direct quantitative estimate of the test's validity is provided by the actual test-criterion correlation corrected for attenuation in the criterion scores but not for attenuation in the test scores.

Sometimes the test *is* the job, in which case the question of validity reduces simply to the question of reliability. In other cases the job itself is of such a type that the ultimate criterion performances cannot be observed or judged or scored. We can watch a person read, but we cannot observe his comprehension of what he reads. The best possible estimate of reading comprehension is one based on a test performance. In such cases we make the test as relevant as we can on the basis of logical analysis

and research, but we cannot compute its validity, though we can and should investigate its reliability.

The definition of relevance

The relevance of a test to a job is entirely a matter of the closeness of agreement between the "true" test scores and the "true" criterion scores. Neither the reliability of the test nor the reliability of the criterion is involved in the concept of relevance. A quantitative estimate of a test's relevance to a specified task is given by the test-criterion correlation corrected for attenuation in both the test scores and the criterion scores. This correlation indicates the closeness of agreement between what the test measures and what the criterion measures, without reference to how reliably either the test scores or the criterion scores measure whatever they do measure.

Predictive power

We have noted that the validity of a test is not a function of the reliability of the particular set of criterion scores which is used in determining its validity. However, there are some criteria which, though imperfectly reliable, are important in their own right rather than merely as samples from a hypothetically infinite universe of behavior of the same type. These criteria usually carry their own labels of success or failure or achievement, and tests are used to predict success or failure, or the final levels of achievement. Some such criteria are of the nature of terminal events: graduation or nongraduation from high school or college, discharge or nondischarge from a mental institution, and retention or nonretention on a job after a tryout period. Others provide quantitative evaluations of achievement or proficiency over some specified period, or of the time needed to reach some specified level of accomplishment or improvement: freshman grade-point average, months of institutional treatment preceding discharge, and annual merit rating on job performance. The upper limit of the predictability of such a criterion is its own reliability. The predictive power of a test for such a criterion is defined in terms of the raw test-criterion correlation—the correlation between the actual test scores and the actual criterion scores.

Summary

The definitions discussed up to this point may be summarized briefly as follows:

1. A set of *ultimate criterion scores* is a set of unbiased (but not neces-

sarily highly reliable) evaluations of quality of performance on a defined task, made on the basis of a representative sample of observations of task performance of each member of a representative sample of persons from a specified population.

2. The *predictive power* of a test is the correlation between the raw test scores and the raw criterion scores.

3. The *validity* of a test is an estimate of the correlation between the raw test scores and the "true" (that is, perfectly reliable) criterion scores.

4. The *relevance* of a test is an estimate of the correlation between the "true" test scores and the "true" criterion scores.

Since a test can be used as an indicator or predictor of performance in any number of situations, it can have as many validities, relevances, and predictive powers as the number of different criteria with which it is correlated. The formulas and the appropriate experimental designs are discussed in the last section of this chapter.

Fundamental problems

Before we can measure any aspect of human behavior, we must define the function which we intend to measure and the group in which we propose to measure it. The purpose of any set of measures of ability or educational achievement, whether it be a set of test scores or a set of criterion scores, is to assign to the members of some given group numbers which are indicative of their relative performance levels on some particular job, or of the relative degrees to which they have attained some specified educational objective. In rare instances we may be able to measure performance levels or degrees of attainment with reference to a standard external to the group, but in most cases we can measure only the individual differences among its members.

The definition of the function to be measured must specify the acts or operations of which it is composed, the materials acted upon, the situations in which the acts or operations properly take place, the results or products of these acts or operations, and the particular aspects or features of the acts or of their results or products which are to be considered as germane to the function. This is the problem of purpose, which has been discussed at some length in chapter 5.

Next we must arrange to observe a representative sample of the specified acts and operations, or of their results or products, in a representative sample of the defined group, and to score or evaluate the behavior or its results or products without systematic bias. This is the problem of the criterion. To the degree that the function cannot be defined in terms of

operations or results or products which are open to external observation (either directly or indirectly), and to the degree that the specified aspects of these operations or results or products cannot be observed and evaluated without bias (that is, independently of all other types of behavior) and with some substantial reliability, the problem of the criterion must remain insoluble.

We may desire to estimate the probable quality of the criterion performances of individuals for whom direct observation and evaluation of such performances are impossible or unfeasible. We therefore devise some substitute performance consisting of more or less different materials presented in more or less different situations. These substitute performances are usually so designed that their quality or the quality of their results or products can be evaluated unambiguously and reliably and easily. This is the usual problem of test construction. A second test construction problem arises when the criterion behavior as defined is intrinsically unobservable or impossible to evaluate (as in the case of reading comprehension).

Finally, we should investigate the relation between evaluations of the criterion operations or their results or products and evaluations of the test operations or their results or products. This is the problem of validity.

The Criterion

Defined action series

Human behavior may be thought of as consisting of actions. In studying the behavior, we classify the actions upon the basis of who performs them, the processes carried out, the equipment or other materials used, the things acted upon, the changes brought about by the processes, the times at which the actions are carried out, and various other bases. Among all acts, for example, all acts of giving may be singled out.

It must be supposed that "giving" can be defined, so that we can attain a reasonable degree of agreement as to whether or not any act shall be called giving, and thus put it in this class if it belongs there.

Among all acts of giving (in which the process is the same) a smaller group of acts may be classified by similarity of the content upon which the process is carried out, as, for example, all acts of giving money. Among the acts of giving money are acts of giving money to a church. Here the classification is upon the object of the giving. Among these acts are those of giving money to a church on a Sunday at the time of the offertory collection. Here the feature common to the acts is a relative position in a recurrent time cycle. Among all these are the acts carried out by Mr. John

Doe. Here the classification is upon the basis of the performer of the act.

Within this series of Mr. John Doe's acts of giving money to a church at the time of the offertory on Sunday mornings, there may be variations in the amount given, the exact time of day at which the act is performed, the time consumed by the act, the form in which the money is given, and the details of the manner of giving.

Criterion scores

Some of the variations are commonly regarded as much more important than others. For example, variations in amounts given may be of much greater interest than variations in the time of day at which the offertory is collected, or in the expressions on Mr. Doe's face from one occasion to another, or in the combination of coins making up the amount given. If we have reasons for studying the amounts given regardless of the combination of coins, the time of day, or the facial expression at the time, we may then disregard these latter variations and describe the entire series by reporting simply the amounts given. If the amount of money at each offertory is taken as the score for that action, we say we "score" the act by the amount given. The score might also be merely "yes" or "no," to report that money was given or not given on a certain occasion in the series. If the score is the amount given, a score of zero may be assigned when no money is given, and the amount may be reported whenever any money is given. If we are interested in Mr. Doe's generosity rather than in the amounts accruing to the church, we may decide to score each act by dividing the amount given by his weekly earnings.

When a series of actions is scored, some summary statement of the individual action scores is taken as the score for the series. If the actions are scored numerically, as in terms of the amount of money given, or the amount given divided by the weekly earnings, then the summary score for the series is commonly the sum of the separate action scores or some type of average. If the actions are scored "all-or-none," as "gave" and "didn't give," then the summary figure for the series would be a number stating the number or proportion of "gave" scores among the total number of scores or actions: the number or proportion of Sundays on which Mr. Doe gave something at the offertory.

Whether the score for the series is a total amount score, or an average amount score, or a total- or average-amount-divided-by-weekly-earnings score, or a statement of the number or proportion of "yeses" or "gaves," the summary score for the series is called the criterion score for the series of actions. It is explicitly a score for the giving practices of Mr. Doe as regards amounts of money given or amounts given in relation to resources,

or frequency of giving, to a church at Sunday morning offertory. It is a measure of one aspect of his giving actions.

We are not ordinarily concerned with such narrowly defined series of actions as the one just discussed, and the actions in which we are interested are commonly not amenable to such easy scoring as in the case of amounts given.

When Alice Jones is in the company of an animal, we may be able to agree upon whether or not a given action of hers is one of kindness to the animal. Even in the case of a kind act, we may fail to agree as to whether it is very kind or just barely kind, and we may fail even more completely to agree upon any numerical scale of kindness for the act. We are failing to some degree in the task of scoring the action for kindness. This makes our task of studying such actions less useful, but if we can agree concerning one action that it shall be called kind, and concerning another that it shall be called unkind, and concerning another that it shall be called neither, then each act may be scored as kind or unkind or irrelevant, and a summary statistic concerning the series may be derived: the ratio of "kinds" to "kinds plus unkinds." This score for the series is the criterion score.

Multiple attributes

The scoring of criterion behavior so as to have the score correspond to the goodness of the behavior is, next to making the observations at all, one of the most difficult problems in the determination of the criterion score. Frequently the scoring varies with the practical value of the performance. In combining a speed score and an accuracy score for a typing performance so as to get a single goodness score, we may in one situation (as, for example, in the preparation of typescript for photo-offset publication) place most of the emphasis upon accuracy and little upon speed. In another situation (as for example, in typing the rough draft of something to be edited) we may want mostly speed without necessarily any great neatness or accuracy. In either case we try to combine the two scores in such a way that the composite typing skill score is higher when either speed or accuracy is high, and higher still when both speed and accuracy are high. By assigning the weighted sum of the accuracy score and the speed score to the typist as her skill score, we recognize that we ascribe a higher degree of skill to a performer who takes less time to perform a task at a given accuracy level, as well as to one who performs the task more accurately in the same length of time. In assigning the weights, we must ordinarily resort to value-judgments concerning the relative im-

portance of speed and accuracy in connection with the particular job.

So also with reading performance. We may desire speed of reading in a college student in order that he can cover his assignments, but we also want him to understand what he reads, so that whenever he is asked questions about his reading, he can give satisfactory replies.

The difficulty in deciding how to combine speed and accuracy, quantity and quality, reading speed and comprehension, and other such multiple scores on single performances has led in many instances to the preservation of the duality. Typing speed and accuracy are commonly both reported, and so also are reading speed and comprehension.

Complex behavior series

The pilot of an airliner may fly the craft from New York to Nashville nonstop. If he performs this service on alternate days, there is a series of performances. We may be able to describe certain features of each performance as desirable, such as planning ahead, working well with his copilot, keeping track of power settings, ground speed, fuel supply, and the like, and following the course which puts the aircraft on the ground at Nashville on time. The same features may be observable on the return trip to New York. If we can agree upon some method of combining our evaluations of the manner in which he handles the various features of the job, such that a high total attaches to a generally desirable performance, then we can score the series of performances and refer to the sum or average of the scores as the pilot's proficiency score for the series.

We should note, however, that this kind of a criterion score is different from those of John Doe and Alice Jones and the typist. Mr. Doe's giving score was expressed in dollars and cents; Alice's kindness score was based on many kinds of actions, but every one of them could be scored, insofar as its kindness was concerned, by reference to a single standard, namely, the resultant benefit or injury to the animal. The standard, moreover, though subjective, is one which can be defined with fair accuracy. If the observer merely *reports* one of Alice's actions toward an animal, and the immediately following actions of the animal, almost any sensible person could assign a kindness score to Alice's action. Acts of kindness are essentially acts of single class.

The typist's speed, like Mr. Doe's giving, yields an objective score. Her accuracy also yields an objective score if we agree that, for all practical purposes, one error is as bad as another. Her combined skill score is also objective; it is determined by some agreed-upon formula from her speed and accuracy scores, and the undesirability of a given error does not de-

pend upon the typing speed at the time that error was committed.

The airline pilot's performance is much more complex than either John Doe's giving or Alice's kindness or the typist's skill. Each action must be referred to a different standard of performance, and must be *evaluated* by the observer in terms of the total context of the circumstances under which it occurs. Kicking hard on the right rudder pedal of an airplane may indicate high, medium, or low flying proficiency, depending on a host of attendant circumstances. In the case of flight performance, furthermore, there is not even a single standard of judgment for the results achieved, let alone for the actions whereby the results are brought about. The "good" or proficient pilot is one who completes a high percentage of his flights, but turns back whenever it is dangerous to continue; who conserves fuel, but uses it lavishly to gain sufficient altitude when crossing mountains; and who stays on the course and arrives on schedule, but flies far out of his path to avoid storm areas. Thus, it is clear that there is no single standard by which to describe goodness or badness of over-all performance. The pilot's proficiency score is entirely subjective. It is an evaluation. There is even some question as to the propriety of using the term "measurement" at all to designate such a score. Two pilots could conceivably obtain the same score on the basis of action patterns having no single feature in common.

Defining the criterion series

Alice's kindness to animals is exhibited only when she is in the presence of animals. Mr. Doe's giving opportunities were scheduled on successive Sundays. The aircraft pilot's flights were also scheduled. In the case of the Sunday giving and the flights on alternate days, we envision a more or less regular schedule of the situations in which we score the subject's behavior. We do not anticipate any regular schedule for Alice's encounters with animals, and what we regard as desirable is that *whenever* Alice is in the presence of animals her behavior shall be of the sort scored "kind." This "whenever" conception of the criterion series is very common and useful in educational work. We say, for example, that "whenever" a person is presented with the numbers 9 and 6 under conditions in which they ought to be added, his response should be 15. We do not anticipate any scheduled series of occasions on which this pair of numbers shall be presented to him, but we still consider the criterion series to consist of his acts "whenever" 9 and 6 are presented to him in an addition situation.

Criterion series behavior need not even extend throughout the lifetime of the individual. If young automobile drivers are to become mature and responsible adults, they must survive the risks of their youthful driving,

and we may properly be interested in the driving skills and habits of youngsters as youngsters.

In the case of criterion series behavior, whether or not it extends throughout the life span of the individual, we think in terms of an indefinitely long series of acts which the individual *might* perform. We think of the young driver's proficiency and skill "whenever" he may drive. And if we observe a sample of his driving behavior, we think of it as a sample drawn from a very large amount of such behavior, even though the number of times he actually drives a car during, say, his senior year in high school, is relatively small. The indefinitely large number of times that we think of as the universe of behavior from which our sample observations were drawn is provided by the "whenever" conception. "Whenever" he might have driven a car, had he done so, we would have been interested in his performance. We observed only some fraction of the times he actually did drive.

In the "whenever" situation we actually define the indefinitely long series in terms of the sample we observed. If we have watched him driving a car on a number of occasions which we regard as an acceptable sample of his driving behavior, we may say that the criterion behavior in which we are interested is "such" behavior—the kind we observed. The sample should ideally be selected to be representative of *all* his driving behavior, and not merely of his driving behavior when there are adults in the car; this leads to difficulties which are discussed later. In any event, the definition of the universe of behavior as behavior "such as" the kind observed automatically makes the observed sample a *random* sample from the criterion series, and we use the formulas for random sampling, including reliability formulas, in our evaluation of the observations. Thus, even if we had a score for every automobile-driving performance which a high school student carried out during his senior year, we would still regard it as a sample. It is a sample from the much larger amount of driving that he might have done had he done it.

Criterion sample scores

As a practical matter we cannot wait until the end of Mr. Doe's lifetime to score our observations on his givings, nor to the end of Alice's lifetime to evaluate her kindness to animals. As a practical matter we cannot hope even to make observations on all the occasions on which Alice encounters animals. Nor can we as a practical matter ride with the airline pilot on every trip. Even if we *could* do these things, we would be able to observe only the behavior when it *did* occur, rather than on all the occasions on which it *might* have occurred. We therefore necessarily

take recourse to sampling from the series. By assigning a score to each act in the sample, we arrive at a sample score. If the sample consists of a random selection of acts from a singly classified criterion series, the sample score will deviate from the true series score in accordance with the laws of random sampling. If the criterion series is complex—that is, if the acts which make up the series fall in several subclasses—the sample must be representative. It must consist of observations of a random sample of acts from *each* subclass, the numbers of observations of the several types of subseries behavior being proportional to the relative frequencies of such acts in the total series.

If the laws of random sampling are to apply, it is also necessary that the observational errors and the errors of evaluation be unrelated to the quality of the behavior observed. In this case the observations and evaluations are said to be unbiased. If the acts observed are a random or representative set of acts from the defined series, and if the observations and evaluations are unbiased, the resulting sample scores are termed "criterion scores." They are criterion scores because they differ only randomly, and not systematically, from the true series scores. *Hence a set of criterion scores has, by definition, perfect relevance,* though, of course, imperfect reliability.

Criterion scores and test scores

Ideally we should not use the term criterion scores for any measures which fail to meet the requirements of random or representative selection of acts from the criterion series, and unbiased observation and evaluation. The relationship between scores derived otherwise and the hypothetical scores from the complete series is not given by sampling theory, but must be determined by experiment if possible, just as though the observations came from outside the series. In educational measurement a great many of our observations do come from outside the criterion series, and scores from such observations are distinguished from criterion scores by the name "test scores." Test scores may also come from non-random or non-representative samples from the criterion series, or from biased observations or evaluations, in which case it seems best still to call them test scores, as distinct from criterion scores, which come from *random* or *representative* samples from the criterion series, and *unbiased* observations and evaluations. If the behavior observed is fairly similar to the criterion behavior, we term such tests "related behavior" tests (chapter 5); if the test behavior is a biased sample of the criterion behavior, or a representative sample of such behavior scored on the basis of biased observations or evaluations, the test is termed an "identical elements" test.

In scoring the acts in a criterion series, we are bound by the desire to obtain a score which represents the general goodness of the performance in some sense. It would not suit us to give Alice a "2" for every act of kindness, a "1" for every act of unkindness, and a "0" for every irrelevant act if we are computing her kindness score. Since the criterion series contains acts which we evaluate for their own sakes, we wish a score which describes the total goodness or desirability or aggregate value of these acts.

In a test series, however, we are not ordinarily concerned with the performances per se, but are interested in the test scores as indicators of the criterion scores. We therefore score the acts in the test series in whatever way will maximize the correspondence between the test scores and the criterion scores. If it is known by extensive experimentation that people who punctuate most correctly and who employ the least incorrect punctuation in the writing they do in the German language are those who, when given a test passage of unpunctuated German, put in the largest amount of correct punctuation, and, in addition the largest amount of incorrect punctuation, then we properly score this *test* by giving some positive credit for incorrect punctuation as well as more credit for correct punctuation, even though incorrect punctuation is undesirable in the *criterion* series. We must be careful here to distinguish between "predictor" tests and "instructional" tests. In constructing the latter we are bound by the same considerations which apply in criterion development.

The combination of different part-scores—such as speed and accuracy, "kind" and "unkind," or quantity and quality—into a single composite score is commonly, in the case of the criterion, carried out with the idea of arriving at a single score which when high indicates good behavior and when low indicates poor behavior, regardless of any correspondence with other behavior outside the criterion series. In the case of predictor test behavior, the combining of the part scores is done so as to maximize the correspondence between the test scores and criterion scores. Thus, the combining of the part scores is done with reference to behavior outside the test series, namely the criterion behavior.

Working criteria

A biased sample from the criterion series is still a direct indicator of some part of the criterion behavior. A test is often an indirect indicator of such behavior. We are forced in many cases to estimate the validity of an indirect indicator as best we can by finding its correlation with a biased direct indicator of the behavior, in situations wherein it is impossible to obtain a strictly random or representative criterion sample, or to score the sample without bias. In such situations we may judge that the bias is not

too great, or that we may accept a nonrepresentative sample of criterion behavior as a sufficient approximation to a representative sample for the purpose at hand. We may also have reason to judge, from the nature of the sampling or bias and the nature of the test, that the test might correlate substantially with the whole criterion series, but could not correlate to any appreciable degree with the nonrepresentative element or bias per se. In either case we will judge that whatever the correlation between the test scores and the imperfect criterion sample scores turns out to be, the correlation between the test scores and a set of representative and unbiased criterion scores would be quite similar. In this case we would term the nonrepresentative or biased sample from the criterion series a "working criterion." A representative and unbiased sample from the same series would be an "ultimate criterion." When the term "criterion" is used, without any qualifier, an ultimate criterion is always implied. But when we are using a working criterion, we reserve the term "test" for the indirect or less direct indicator of the behavior. The term "criterion" is commonly applied to any set of measures or observations that may in fact be used as a standard in estimating the validity of a test. Unless it is actually an ultimate criterion, however, we must be very careful to include a qualifier. Working criteria do not possess perfect relevance, and the assumption of "practical identity" is beset with a variety of dangers unless the relevance of the working criterion has actually been determined on the basis of its correlation with the ultimate criterion. A set of nonrepresentative or biased criterion scores may well be less relevant to the ultimate criterion than are a set of scores on a carefully worked-out test.

Intermediate criteria

Often we are unable to obtain any sample at all of the actual actions which we wish to evaluate or measure. They may lie too far in the future, they may be inaccessible to systematic observation, or they may be of such nature that unbiased observation or evaluation is impossible. In some such cases we may resort to predicting events which according to our best judgment *ought* to be correlated with the behavior we want to evaluate. In other cases we may be able to obtain samples (often somewhat nonrepresentative or biased) of some of the criterion subseries but not of others. Such criterion samples are clearly imperfect; we cannot assume that they are "approximately valid" or "sufficiently valid" for the purpose at issue. Hence, they are not even working criteria. They are last resorts, but in very many cases they are necessary last resorts. We term them "intermediate criteria." If a test correlates substantially with an inter-

mediate criterion, we can infer only that it will probably correlate to some extent appreciably greater than zero with the ultimate criterion. If it correlates zero with the intermediate criterion, it can correlate substantially only with those elements of the ultimate criterion which are independent of the intermediate criterion.

A somewhat different type of intermediate criterion may be used in cases where the ultimate criterion behavior (usually job behavior) can be exhibited only by those who have previously met prescribed standards in areas of behavior whose relations to the ultimate criterion behavior are likely to be at best imperfect. Before we predict success in medical practice, we must predict graduation from college, graduation from medical school, and passing of a state medical board examination.

Suppose we are unable to make direct observations of "success in engineering," or even to define it accurately and unambiguously, but still wish to construct an engineering aptitude test for use in counseling high school seniors and college freshmen. First we may predict success in mathematics, if only because failure in mathematics bars the student from further professional study in engineering colleges. Then we may predict graduation or nongraduation, or final grade-point average in the engineering college. We know that success in the engineering profession is not the same thing as our "criterion" (grade-point average) of success in the engineering college, but we have reason to believe that there is some fairly substantial relation between these two types of success. We should not rest content at this point, however. We should correlate both the test scores and the grade-point averages with, say, salaries at age thirty-three (or ten years after graduation), election to honorary engineering societies, and any other partial indices of engineering success that may become available. If either or both of our early scores (test score or grade-point average) correlates substantially with *several different* biased indicators of ultimate success, we shall believe we have increasing evidence that it would also correlate substantially with an unbiased indicator if such could be found.

There are many more problems concerning intermediate criteria which are of greater importance in industrial and military psychology than in education. These problems are discussed in the references given at the end of this chapter.

Recording and judging

Some performances carry their own tags of excellence. A sum is either right or wrong. Mr. Doe's giving is a matter of dollars and cents. The typist's

speed score can be computed in strokes per minute, and we can determine her accuracy score in terms of errors per hundred lines typed (after some not insuperable difficulty in setting up a schedule which will define all types of errors—or nearly all types). Speed of reading can also be defined in terms of words per minute.

In the case of many other performances the excellence or badness is inherent in the acts themselves rather than in their products or their frequency. The defined aspect of each act which is to be scored must be observed at the time the act occurs, and recorded then or later. In such cases observational errors and errors of recording are bound to occur at least occasionally, and they often occur with disconcerting frequency. In still other cases, the observer does more than merely record the pertinent aspect of each act as he observes it; he must also *judge* whether a given act is good or bad or neither, and if good or bad, perhaps how good or how bad. If we define a "kind" act toward an animal as one beneficial to the animal, and an "unkind" act as one injurious to the animal, the observer of Alice's acts must know enough about animals to judge whether a given act is beneficial or injurious or neither. (Acts of this last class may be judged to be irrelevant to the kindness-unkindness continuum, and hence not scored at all rather than being scored "neutral.") There will inevitably be some lack of agreement among equally competent observers as to whether or not certain acts do or do not fall on the "kindness" continuum, and as to whether some acts should be classified as "kind" or "unkind" (beneficial or injurious). If these lacks of agreement are simple matters of imperfect human perception and judgment, there is little loss; the criterion scores will be essentially relevant but subject to some observer unreliability. However, one observer may judge an act in terms of one concept of what is beneficial and what is injurious, while another may judge the same act by reference to a different concept. Complete agreement can never be attained in practice. The "kindness" of an act is not a function of the act alone; it is a function of both the actor's act and the observer's act of judgment.

In the case of the acts involved in piloting an airplane, the situation is even more complex. The concept of excellence must change with each act observed; there is no single standard comparable to "beneficial" versus "injurious." Hence the competence of the observer is just as crucial to the validity of the pilot's "excellence" score as is the nature of the acts performed by the pilot. We should do everything we can to improve the objectivity of the observations and judgments, but we must recognize that no close approach to complete success is possible. We should therefore, ideally, determine the competence of the observer first, by observing his

acts of observation and scoring them. But before doing this, we should determine the competence of the observer of the acts of observation in the same manner. This obviously leads to an infinite regression, and the problem is ultimately insoluble in any absolute sense. There is no finite set of operations by which one single act of piloting may be evaluated in terms of its excellence; an infinite number of preliminary agreements must be reached before we can agree that the record of a given observation is to be termed a valid record of the excellence of the act observed.

Wherever it is necessary to impose an observer's interpretation or judgment between the act observed and the record of its excellence, we have a subjective observation. Subjective observations never possess perfect relevance, because the standards and operations by which the observer arrives at an interpretation or evaluation cannot be specified completely. Hence, complete prior agreement concerning these standards and operations is impossible.

In the final analysis, of course, all observations are subjective. But operations such as counting, adding, recording the times at which a given series of acts by the person observed began and ended, and the like, possess such a high degree of objectivity that they are commonly termed "objective." Even the observations of Alice's acts of kindness and unkindness are somewhat "objective," though a certain minimum of interpretation (and hence of subjectivity as we are using the term here) is necessary. We shall term an observation "objective" whenever there is a generally agreed-upon standard which can be communicated from one observer to another with little or no ambiguity, and when the observer is not required to interpret or judge, but only to record his observations.

Sometimes we can agree upon a class of persons who ought to be competent judges or interpreters of the acts to be observed, or who as a class are presumably more competent than members of any other class. The class of commercial airline pilots might be agreed upon as such a class, for the purpose of observing and scoring flying excellence or proficiency. We might select a random sample of such pilots, and have each of them observe and judge independently the excellence of the acts of piloting of our particular pilot. Their average judgment should be more nearly valid (both as to relevance and as to reliability) than the judgment of any one of them. This is about as near as we can come at present to the achievement of perfect relevance and validity for a series of acts whose excellences must be judged subjectively.

Use of rules

Imperfect rules are better than no rules in narrowing the range of

disagreement. The rule for defining "kind" and "unkind" acts is a case in point. The pilot's observer may decide before observing that acts which preserve the safety of the passengers are, in general, ten times as important as are acts which tend to keep him on schedule, and fifty times as important as acts which save gasoline. In scoring a piece of writing, we may decide that the omission of a verb is, on the whole, fifteen times as bad as splitting an infinitive. Such rules cannot be applied mechanically, but if several observers agree upon them as general guides to judgment, their use will reduce the variability of the judgments to some extent.

Use of subscales

In the case of complex series of acts, such as those comprised in piloting an airplane, it is often advantageous to score the acts in each of a number of subseries separately. Thus, the observer might consider take-off acts in one group, acts of managing the plane in flight in another, acts of navigation in a third, acts in response to unusual flight conditions in a fourth, acts of landing in a fifth, etc. In this case it might not be necessary to combine the various subseries scores on the basis of direct judgments. A number of observers of different pilots might each assign several subseries scores to the pilot whose behavior he observed, and also assign this pilot an over-all flight performance score. The weights to be assigned to the subseries scores could be determined by multiple correlation procedures based on their intercorrelations and their correlations with the over-all scores. Each pilot's composite score would then be the weighted average of his subseries scores. *These composite scores would probably be considerably more reliable, and hence more valid, than the over-all scores.* The over-all score assigned to each pilot is assumed to possess greater relevance but less reliability than any combination of the subseries scores weighted by direct judgment. Similarly we can derive sensible weights for combining speed and accuracy scores on a typing performance test by asking supervisors to rate typists on an over-all scale of proficiency, and then observing separately their speed and accuracy in actual typing performances. So long as we stick to a set of jobs which are judged to require similar ratios of speed to accuracy, this method is likely to be better than the usual alternatives. It is likely that judgments of over-all excellence of performance will be more valid in many cases than will judgments concerning such matters as the number of strokes that should be subtracted for each error.

Use of behavior check lists and rating scales

One judge may be struck by one aspect of an act or series of acts, while

another judge may be more impressed by some other aspect. Quite apart from variations in subjective standards, we must take account of differences in the directing of attention, and of differences in selective memory between one judge or observer and another. In practice it is seldom feasible to record a judgment of the quality of every act as it occurs; if the acts come in rapid sequence, the observer might fail to note some of them because he was busy evaluating and recording others. The observer of the pilot's flight performance would probably find himself in this situation from time to time. Hence, the summary record must often be made some time after the occurrence of the acts observed.

In such situations the use of behavior check lists and rating scales is recommended for several reasons. In the first place agreements concerning what acts are to be observed (and even what acts are to be disregarded) become a matter of written record. Check-list items remind the observer of all the more important things he is to look for, thus serving in some degree to control his attention, so that different observers will be more likely to notice the same items. If a check list is long, it will caution the observer against failing to note and evaluate certain acts that might otherwise be overlooked, particularly if certain other acts are conspicuously good or bad.

Rating scales serve to define the subscale allocations of various acts. They also make the judgments of different observers more nearly comparable by forcing the judgments into the same language molds. It is important, of course, that such scales be based on very careful analysis of the complex series into subseries within each of which the acts can be evaluated in terms of a single standard. There must be an appropriate subscale on which to evaluate every act observed. The scale statements must themselves be unambiguous, else the scale will reduce rather than increase the comparability of the evaluations.

When well constructed and rightly used, however, behavior check lists and rating scales help materially in making the postrecordings and judgments of different observers more nearly comparable, in reducing differences in the interpretation of the basic definitions and agreements, and in minimizing the disturbing effects of selective attention and selective recall.

Critical requirements analysis

Many of the important acts included in a complex criterion series may be acts in whose performance individual differences are small. Also if the series includes a very large variety of acts, it may be necessary, as a practical matter, for observers to confine their attention to those of major importance, and not attempt to observe and record and evaluate those whose final weights would in any event be low. In attempting to redefine

a broad ultimate objective in terms of specific acts, therefore, the most practical procedure often consists in identifying only those acts in whose performance substantial individual differences exist, and for which low scores constitute actual failure in the performance of some important part of the larger task indicated by the ultimate objective. Such acts are termed "critical requirements" of the task. Detailed methods for the identification of critical requirements are discussed in recent literature on industrial and military criterion development (*2, 4, 5, 6*).

Logical Problems of Validity

Operational definitions

Generosity is ascribed to the person who gives more frequently or regularly than others, or who gives larger amounts, or amounts which are larger in relation to his income and his other obligations. Kindness is ascribed to the person whose actions are observed to be kind. Typing skill is attributed to the typist whose work is fast and accurate. Flying proficiency is attributed to the pilot whose performances are safe, dependable, comfortable to the passengers, and economical to the company. The actions in each case define the attribute. If we say that John Doe is "generous to the church," we mean simply that his giving score is appreciably above the average. If we say that Alice is "very kind to animals," we mean merely that her kindness score is higher than scores of the great majority of other people. If we call the typist "skillful," we are saying no more than that her speed and accuracy scores are fully adequate to the demands of her job. Observation of the actions precedes the ascription of the attribute. The score for the actions is a numerical summary description of the pertinent aspects of the behavior. The adjective (generous, kind, skillful, proficient), with or without some adverb such as "very" or "moderately," is a verbal summary description of the same aspects of the same acts. When the available data consist only of observations of a single series of actions, or of some particular aspect of these actions, about all we can do with these data is to describe them. For some purposes we can describe them most usefully by summarizing them. A numerical summary is often but not necessarily more useful than a verbal summary. The single series, however, does not cause or explain or account for either itself or any summary description of itself. We cannot say that Alice's kindness explains or causes or accounts for her kind acts so long as the acts themselves are the only evidence of the kindness, since in this case the "kindness" is merely a verbal summary description of the actions. In communicating summary descriptions, it is advisable to avoid the use

of expressions which in other situations often have causal or explanatory implications.

"Generosity," "kindness," "skill," "proficiency," and the like are words representing abstract concepts. When we ascribe no more meaning to such a term or concept than the common features of the acts summarized by the word, the term and concept are said to be "operationally defined." So long as we stick to operational definitions in discussing and thinking about the abstract concepts used in educational measurements, we will be on safe ground. The reader must be warned that it is often very difficult to talk and think in this fashion, but the rewards in the clarity of both ideas and their communication are well worth the effort required.

Let us say that "ability" is nothing but a term summarizing the behavior of persons whose actions within some defined series are characteristically successful actions; and let us apply this definition to the case of "mental ability." We begin by agreeing whether any given act is to be classed as mental or not. We agree upon a scoring scheme which attaches a high score to the act when the performance is good in some sense, and a low score when it is not good. A series or sample of such acts yields a series of criterion scores, whose sum or average represents the average goodness throughout the series or sample. The performer whose average score is high is termed mentally able. We must not say that his high score is due to his high ability, but if anything the reverse. We say he has high ability because his performance has yielded a high criterion score. His "ability" is simply a summary statement concerning his actions.

We may define the general class of mental acts directly, and then define acts of reasoning, comprehension, judgment, etc., as subclasses of the class of mental acts. On the other hand, we may feel ourselves to be on surer ground if we define a number of subclasses first, and investigate the relations among the subclass summary scores before defining the larger class. Since "general mental ability" consists of whatever acts we agree to employ in defining it, we may agree not to reach agreement on such acts until we know the interrelations among the subclass summary scores. We are pretty sure to include acts of reasoning and acts of comprehension and acts of judgment in the general category of the acts to be used in defining "general mental ability." Whether we will include, say, acts of suspending judgment, is perhaps another matter. It is quite possible that we could agree to include such acts if, but only if, the correlations between sets of summary scores for suspending judgment and those of the same people for reasoning and comprehension and judgment were positive and substantial. We might even agree to define "general mental ability" in some very indirect fashion: as some mathematical function, for example, of the

summary scores from a number of subseries. The particular subseries to be included, as well as the combining function to be used, might be chosen only after careful consideration of the results of factor analyses of large numbers of subseries scores based on types of acts which have been proposed but not yet generally accepted as mental acts. This indirect method of reaching agreement may succeed in some instances where all direct methods fail.

Explanation

Nothing explains itself, and no action is caused by its own occurrence, hence our insistence that the phraseology of a verbal summary statement concerning a set of observations must not be interpreted as an explanation or causal inference. An explanation consists in showing that there is a relationship between the performances to be explained and *other* performances, on the basis of which the given performance could have been predicted from a knowledge of the other performances. The other performances may be other performances of the same individual, but more commonly they are the performances of other people. Such relationships, when well established by observation, become explanatory concepts.

Mr. Doe's contributions to the church may be quite regular in the sense that he makes some contribution every Sunday, but they may still show irregular fluctuations in amount. If Mr. Doe is a commission salesman, we may discover that his weekly givings are almost always a given fraction of his weekly earnings. Knowing his earnings for a given week, we can make a better prediction as to how much he will give next Sunday than we could have made on the basis of complete knowledge of all his past givings. The fluctuations in his weekly income explain the fluctuations in his Sunday givings.

Suppose Alice Jones has lived in a city all her life, and that her actions in the presence of animals have been limited to actions in the presence of dogs, cats, and horses. We observe the actions for five years and rate them as "kind" or "unkind," with the following results:

Year of Observation	Percentage of Kind Actions toward			Mean	Range
	Dogs	Cats	Horses		
First	51	66	56	*58*	15
Second	63	59	49	*57*	14
Third	50	65	58	*58*	15
Fourth	80	75	70	*75*	10
Fifth	88	83	86	*86*	5

We note that during the fourth and fifth years the means have risen and the ranges have decreased. We may summarize the first fact by saying that Alice has become kinder, and the second by saying that her kindness

has also become habitual. Habituation is an observable aspect of sets of actions. Actions of a certain class are said to be habitual when they are observed to be consistent. The process of habit formation is evidenced by a progressive decrease in variability, often but not necessarily accompanied by a progressive increase or decrease in the accompanying average. These processes are observed in many people and in many classes of actions. The term "habit" is, therefore, a general explanatory concept. From the evidence at hand and our additional general knowledge of habits, we infer that Alice's habit of kindness to animals will exhibit the characteristics common to habitual behavior. Among such characteristics are persistence or stability, and a tendency toward generalization.

Alice now moves to the country, and later observations show that during her first year on the farm her kindness-toward-cows percentage is 85. This score is consistent with her last previous scores for kindness toward dogs, cats, and horses, and entirely inconsistent with any of her scores for the first three years of observation. The new score (as a summary of the actions) is explained by the habit.

We shall not enter here into a discussion of causal logic. Whether or not Alice's habit of kindness to animals (or to domestic animals) "caused" her to be kind to cows, her previous actions toward other species, interpreted in terms of the concept and characteristics of "habit," did explain or account for her actions toward cows. These actions could have been predicted before they were known to have occurred.

If the airplane pilot is consistently poor in navigation, but excellent in take-off performance, landing performance, and airplane management performance, we might give him the XYZ Mathematics Test. If he makes a low score on this test, we shall suggest that his poor navigation performance is accounted for, at least in part, by his demonstrated lack of one type of mathematical ability, as evidenced by his low score on the XYZ Test. We *assume* that there is a relationship between XYZ Test ability and navigating ability because we have observed that in navigation, the pilot has been making computations more or less comparable to those called for in the test situation. The score on the XYZ Test accounts in part for the navigation performance, rather than the navigation performance accounting in part for the test score, because the test measures a much narrower function; that is, a "trait" (see discussion under "Tests and Traits" in a following section, page 647).

Measurement, appraisal, and prediction

A sample score is an arbitrary summary of some observed aspect of a set of actions from a defined series. This is equally true of test series and

criterion series. An act is included in a series because it "belongs" to a certain "class" of acts. The definition of the class is a matter of agreement; that is, of judgment. So also are the operations by which one act is included in the class while another is not. The basis for the definition-judgments and the inclusion-or-exclusion judgments (which represent interpretations of the definition) is usually a set of prior value-judgments. Acts are included or excluded, not because of any intrinsic relationships among the acts as acts, but because they do or do not bear upon some desired outcome or outcomes.

We desire the airline pilot to avoid accidents, to stay on schedule, to maintain the comfort of the passengers while in flight, to avoid undue wear and tear on the plane, to avoid overconsumption of gasoline and oil, etc. These are related values only because they are all related to flying an airplane. The pilot's acts of taking off, managing the plane while in flight, navigating, changing his course in response to unusual weather conditions, landing, and the like, "belong" to the flight proficiency series because they all bear upon one or more of the desired outcomes. The acts may be related among themselves or they may not. The performance score is an *appraisal* of the aggregate value of the acts in producing the desired outcomes.

The term "measurement" is somewhat more restrictive. Unless the item or observation scores are themselves related, in the sense that they vary consistently rather than independently, the summary score is an appraisal but not a measurement. Measurement requires a more or less linear score continuum and a set of more or less equal score units. The linearity of the continuum need not be perfect, and the score units need not be exactly equal. But the continuum must at least exist, and every unit (item or observation score) must have some relation to it. This means that when we score a set of test items or criterion observations for each of the individuals of an appropriate group, every item score must exhibit some significant positive correlation with every other (negative correlation if the scoring weights are of opposite sign); or, at the very least, every item score must exhibit a significant correlation of proper sign with the total score. By an "appropriate group" we mean a group for all of whose members the summary scores are intended to be interpreted as measurements. Note that it is only the scores, not the acts scored, which must possess the required properties.

Sometimes we can usefully limit our appraisals of acts to aspects which can be scored in terms of "external" units. Mr. Doe's Sunday givings are scored in dollars and cents. The typist's speed is scored in terms of strokes

per minute. Speed of reading is scored in words per minute. In these cases the inter-item correlations are an irrelevant consideration. One dollar has the same value to the church as any other. The quality of the typist's strokes and the comprehension of the reader are irrelevant to the speed scores, per se, though in each case the correlation between speed and quality may be substantial. Such scores are essentially metric measurements. The score units as defined are equal, regardless of the equality or inequality of the values which underlie them or of the values to which they lead.

In other cases we try to score some quality of the acts themselves rather than merely recording or counting them and referring the records or counts to an external scale such as the scale of dollars or minutes. We may still employ counting, but we count something about the acts which is not uniform by definition. The child's addition score is the number of right answers he produces when he attempts to add the 100 basic digit pairs. The typist's error score is the number of errors per hundred words (or per 500 strokes) typed. The reader's comprehension score is the number of questions about the passage which he answers correctly. Alice's kindness score is the percentage of kind acts in the kind-unkind series or sample. In these cases the inter-item correlations are relevant and important. We must assure ourselves that missing one combination is equivalent to missing another, or that answering one question right is equivalent to answering another one right, or that one act toward animals scored "kind" is equivalent to another. The particular sense in which they must be equivalent is that each one of them must be an equivalent indicator of the total summary score. Ideally each item score must have the same validity as every other, as an indicator of the summary score, and these validities must all be above zero. In practice we are likely to accept evidence that each item score possesses some significant positive validity as a working approximation to the ideal.

Not only must every item score be related to every other, and hence to the total, but the intercorrelations among all the item scores must be more or less equal. No one of the essential conditioning factors of the test or observational situation may be of overriding importance, and neither may any one reaction of the subject to the test or observational conditions. In particular, as the number of items or observations is increased, every one of them must make a progressively smaller contribution to the total score, and the relative importance of each conditioning factor and general reaction of the subjects must decrease. If the child does not understand the zero principle, and if a number of zero problems have been included, his summary score on the arithmetic combinations test will reflect in part the

proportion of problems containing zeros rather than simply the proportion of all addition combinations which he knows. (This argument is, of course, a weak one if every test contains *all* the 100 combinations.) If one key on the typewriter sticks, the typist's error score will reflect mainly the number of occurrences in the copy of the letter on that key. If the student does not understand the reading test directions, his comprehension score will reflect lack of comprehension of the directions rather than lack of comprehension of the passages. Also, if the reading test contains a lot of questions starting out, "The key sentence in paragraph X is . . . ," and if to him a key is simply a gadget used to lock and unlock doors, his score will again fail to reflect his comprehension of these problems. If Alice's acts toward all dogs in the neighborhood except one are kind, but are unkind toward that one, her kindness score will reflect mainly the percentage of encounters with that particular dog. In all these cases, increases in the number of observations or items will not reduce the proportionate importance of the one recurrent disturbing action or condition. In such cases, the summary score is not a measurement, and it is not a valid appraisal either.

When the conditions specified above are met, we have what may, for want of a better term, be called a "quasi-measurement." Quasi-measurements may for many practical purposes be treated as measurements. Their units may be taken as roughly equal, and they may be added, subtracted, averaged, correlated, etc. Their zero points are often indeterminate, however (as are also those of some metric scales) and in such cases they may not be multiplied or divided. We can say that one typist is twice as accurate as another because "no errors" is an observable possibility. We might say that one girl is twice as kind to animals as is another, but the meaning of such a statement would be doubtful at best. We would not say that one reader comprehends twice as much as another, because no one can construct the "easiest possible" comprehension question (one unit of comprehension above zero) and grade the difficulties of all others in more or less equal units upward from that one.

If all the item or observation intercorrelations are significant and of the proper sign, and if no one conditioning factor or action tendency of the persons tested or observed is very important compared to all the rest combined, some sort of measurement exists, but not otherwise. Ideally the system of item or observation intercorrelations should conform to the Spearman two-factor criterion, at least to a close approximation. In this case we have a *homogeneous* test or series: the common-factor structure of every item score or observation score is similar to that of the summary score. How close an approach to complete homogeneity is required before

we can say that a set of summary scores is a set of measurements or of quasi-measurements is at present an unsettled question, with a number of technical ramifications. As a practical matter we are probably on safe ground if we insist only that the score on every item shall possess some validity as an indicator of success on every other item.

The best predictor is one whose items all correlate positively and significantly with the criterion, and zero or as low as possible with one another. Hence, a single test which is an efficient predictor of a complex criterion cannot at the same time be a metric measuring instrument.

Throughout most of this book, the term measurement is applied to any test or criterion summary score which serves a useful purpose. This is all right so long as we are not thereby misled. We must be careful, however, in applying statistical concepts and formulas to such scores. Some of them apply to any kind of scores, while others apply only to scores which satisfy some or all of the essential requirements of metric measurement.

A valid indirect test or battery appraises or predicts some defined set of criterion performances or events. The criterion performances typically draw on a number of more or less unrelated reaction-systems. The criterion summary scores are, therefore, value appraisals rather than metric measurements. If the test or battery is to possess substantial validity or predictive power, some of its items must sample one reaction-system and some another. Ideally the proportion of items sampling each reaction-system should be proportional to the frequency of reactions from that system which are found in the criterion series, and to the importance or value of these reactions in the criterion aggregate. But in practice we wish to use the same tests or batteries for appraising or predicting different criteria. An efficient method of doing this is to construct first a number of short tests, the items of each of which correlate substantially with one another. At the item analysis stage we should use the "criterion of internal consistency." Each of the separate tests will then yield scores which are quasi-measurements. Then in appraising or predicting any given criterion performance, we select tests rather than single items, and weight them by the procedures of multiple regression to obtain maximum validity or predictive power.

Tests and traits

When the item scores of a set of test-item performances correlate substantially and more or less uniformly with one another, the sum of the item scores (the summary score or test score) has been termed a quasi-measurement. It is a quasi-measurement of "whatever," in the reaction-systems of the individuals, is evoked in common by the test items as presented in the

test situation. This "whatever" may be termed a "trait." The existence of the trait is demonstrated by the fact that the item scores possess some considerable degree of homogeneity; that is, they measure in some substantial degree the same thing. We term this "thing" the trait. The greater the homogeneity, the more sharply defined is the trait. A trait may "exist" (as a set of linkages among a set of reaction tendencies) in a single individual, but its existence can be *demonstrated* only when similar sets of linkages are found in other individuals. The linkages are merely conceptual postulates, but the homogeneity of a set of test items can be demonstrated by means of specified operations. We attribute varying amounts of a trait to the individuals of a group whenever it is possible to construct a test consisting of different items, which are responded to with different reactions, but in such a manner that the item scores form an approximately homogeneous set.

Suppose, now, we have a considerable number of quite highly homogeneous tests. We give all of them to the same group of persons, and compute the intercorrelations among the total test scores. We may find that in one or more subsets of tests, the intercorrelations are all high. We say then that the tests of such a subset are measuring "something" in common. That something may also be termed a trait. It is a broader trait than is one defined by a single set of highly homogeneous test items. The methods of factor analysis provide more refined methods of defining such traits. The traits are still defined in terms of the intercorrelations among the measurements. If the tests themselves are not highly homogeneous, the scores yielded by them will not be measurements (or even quasi-measurements), and the intercorrelations among such tests will not lead to adequately precise definitions of broader traits. The "something" common to the scores of one person will not necessarily be the "something" that is common to the corresponding scores of another person.

The methods of factor analysis are superior to the method of simple clusters of highly intercorrelated tests because they remove the need for single-factor homogeneity in the separate tests, which is difficult if not impossible to demonstrate. Suppose an arithmetic reasoning test draws upon three broad traits, say: a verbal trait, a number trait, and a reasoning trait. The items of the test may be homogeneous in the sense that every item of the test measures whatever is measured by the entire test. This might be termed "three-factor" homogeneity. One-factor homogeneous tests are very much harder to construct.

We may expect to find broad cognitive traits, interest traits, attitude traits, personality traits, achievement traits, and various others. Having

found them, we may assign trait scores to individuals, and compute the intercorrelations among the trait scores. We can then factor-analyze the intercorrelations among the trait scores. If clearly defined factors are found, we may term them supertraits. If there is a general factor common to a number of broad cognitive factors, we may choose to call the corresponding supertrait "intelligence." There is evidence indicating that such a supertrait "exists" in the sense defined. There is also evidence that most of what we term "general school achievement" is this same supertrait. The evidence is found in the intercorrelations among tests, some of which are termed "cognitive trait tests" and some of which are termed "school achievement tests."

A number of attempts have been made to identify broad personality traits by means of the usual check list and questionnaire "personality tests." The results of such studies are as yet conflicting and inconclusive. The reason probably lies in the fact that the responses to sets of items of these types have not been shown to possess the property of homogeneity. The same difficulty is found in factor analyses dealing with interest and attitude test scores.

The existence of a trait name does not necessarily imply the existence of a trait, and the existence of different trait names or test names does not prove that the tests measure different traits, nor even that they measure any traits. Ordinary trait names represent value-concepts. The term "trait," as used here, is a measurement concept, being defined by the operations which show that a set of item performances or test performances are to some degree homogeneous. If we want to learn about human traits in any scientific fashion, the only way we can do so is to construct homogeneous tests and then use them in extensive factorial researches.

Ability, capacity, and the AQ

If several individuals have had similar opportunities and similar incentives to develop some sort of measurable ability, and if they do so at greatly different rates, we may say that "whatever" differences among them resulted in the observed differences in rates of development are differences in "capacity." Capacity is thus inferred from different relative rates of progress under presumably uniform conditions of opportunity and incentive. The typical school is an institution par excellence for producing such uniformities. The general cultural environment is certainly a less dependable source of uniform opportunities and incentives for the development of intellectual abilities in general. *Hence a valid school achievement test will reflect differences in scholastic capacities (especially*

among the pupils of one school) to a much greater degree than a general intelligence test can reflect differences in the same or any other intellectual capacities. Insofar as general intellectual capacity affects school achievement, individual differences in this capacity can be inferred better from a valid general school achievement test than they can from a general intelligence test.

The usual interpretations of the accomplishment quotient—AQ—are entirely erroneous, and the actions resulting from such interpretations are often positively pernicious. No general intelligence test can well measure capacity for scholastic achievement (in a single school) as well as that capacity is measured by a general school achievement test of equal general excellence. The AQ is essentially a ratio comparison of two kinds of achievement: scholastic achievement and general intellectual achievement. The corresponding capacities can be inferred, in each case, only on the assumption of equal opportunities and incentives for their development. We will be much nearer the truth if we interpret the AQ as indicating the relative qualities of the educational environment as a developer of educational achievement, and the general cultural environment as a developer of general intellectual achievement. The school itself is the major element of the general environment in the development of educational achievement, and it is only a little less important as a developer of general intellectual achievement. It has been demonstrated that whatever is measured by the Stanford Achievement Test and whatever is measured by the Stanford-Binet intelligence test are about 90 percent the same thing!

Intrinsic invalidity

A proposed ultimate criterion, to be of any use, must be completely relevant to its own purpose. We can define any purpose for education or measurement that we please. If there is widespread agreement as to the worthiness of the purpose, it may be labeled an educational aim or goal. This does not mean that it will become at once a useful ultimate criterion for the validation of tests or for the improvement of the educational program. A useful ultimate criterion demands much more than the statement in abstract terms of a worthy purpose. The statement must imply unambiguously or specify explicitly an area of behavior which is *operationally definable, operationally observable,* and *operationally scorable.*

In the case of some abstractions, we reach no agreement, either directly or indirectly, as to what behavior is involved, or on what basis an act shall be included in or excluded from the series. Such behavior is not profitably studied unless the investigator can at least define his terms to his own satisfaction, and set up inclusion-exclusion rules which are unambiguous

to *him.* And such abstractions—assigned to persons without reference to these persons being performers of acts in a series in which the acts are included or excluded by an accepted rule or set of rules—such abstractions are not useful symbols for types of human behavior. They are intrinsically invalid symbols because they are operationally undefinable.

A criterion series may be more or less definable but not observable. Of such nature are the acts which occur mainly "inside" an individual. We can define after a fashion the acts which constitute such things as frustration, dissociation, anxiety, integration, and the like, but we cannot observe them directly. They are inferred as causal antecedents of observed acts. A single observed act, however, is believed to be the resultant of a great number and variety of these internal acts. In the attempt to get somewhat closer to one-to-one correspondence between observed acts and their presumed "internal" antecedents, clinical psychologists are expending much effort in the development of projective tests, situational tests, and the like. When a great deal is known about the intercorrelations among observed acts, including the acts made in response to projective, situational, and other tests, we may be able to revise the definitions of internal acts, and even to measure them indirectly. Discussion of the necessary procedures is quite beyond the province of the present book. Internal acts are somewhat analogous to the actions attributed to electrons, protons, etc., none of which are directly observable either. The main difference between internal and external acts lies in the much greater complexity of internal acts and of their relations one to another and to overt, observable acts. When and if clinical psychology becomes an exact science, we can be sure that nuclear physics will seem a very simple discipline by comparison.

At present we are on safe grounds only if we list most of the postulated internal acts as operationally unobservable, either directly or by strict inference from objectively observable acts (as in the case of habit, for example). Any postulated series of such acts is, therefore, intrinsically invalid. If we try to use terms defined by such series, these terms will lack meaning. Their use in communication will lead generally to confusion, and action judgments based on them will be fallacious as often as correct.

A criterion series may be both definable and observable, but still impossible to evaluate because no agreement can be reached on how each act in the series is to be scored. It might be relatively easy to count the number of opportunities several individuals had to vote, over a reasonable number of years, and the number of these occasions on which they did vote. But the essential value of voting behavior, as an aspect of "good citizenship," is its quality rather than its quantity. Perhaps A votes only when he feels he knows the candidates and issues well enough to vote

intelligently, while B votes a straight party ticket at every election, and these matters are not open to external observation. The fact of voting or not voting may be observable, but the quality of the voting cannot be evaluated except by reference to these other variables, which are not observable. Another aspect of "good citizenship" might be obedience to the laws; but if we take conviction for a crime as the evidence of disobedience, we will include among the unconvicted not only the law-abiding citizen but also the clever and effective criminal. The *reason* for absence of conviction, which is not observable, is essential to an adequate scoring procedure. If the acts in a definable and observable series cannot be scored except by reference to other acts which are not observable, the series is operationally unscorable and hence intrinsically invalid.

Among the abstractions which we must at present consider intrinsically invalid we find most of the action series that go to make up "worthy home membership," "good citizenship," "democratic attitude," "loyalty," and many of the other ultimate aims of education. On the other hand "command of fundamental processes" does lead to essential agreements. We can fairly well specify the acts performed in appropriate situations by persons to be designated as having such "command," the acts performed in similar situations by persons who are to be labeled as lacking such "command," the materials upon which the acts are to be performed, and the bases upon which the acts are to be classified and scored as successful or unsuccessful. Those educators who insist (and rightly, we believe) that other aims are at least equally important, and in aggregate probably much more important, would advance their cause most rapidly and effectively by setting about the task of specifying the materials, actions, situations, and scoring criteria implied by the abstract terms which denote these other aims. They will find the task difficult but in most cases possible. When they have accomplished it, they will find that teachers will use the materials, set up appropriate school situations, and teach the desired acts. In those few cases where the task turns out to be impossible, the abstract aim must be admitted to be intrinsically invalid.

The Estimation of Validity

Educational objectives

In aggregate, the ultimate criteria for all educational achievement tests are defined by the ultimate aims of education. These ultimate aims, however, are usually stated in terms of broad generalities rather than in terms of explicit rules for the classification and evaluation of observable acts. Most of them, moreover, refer to acts which will occur (or fail to occur)

long after the end of formal schooling, and many of them refer to acts which, when they do occur, are not likely to be observed and scored by those who are directly concerned with the evaluation and improvement of the educational process. In general, therefore, statements of the ultimate aims of education do not lead directly to the development of sets of ultimate criterion scores which can be used in validating either educational achievement tests or educational programs and procedures. Even if long-term follow-up studies were generally feasible, we could not afford to wait twenty or thirty years to find out whether our tests were valid or not; and if we did, the objectives, or at least the methods and organization of the educational program (and the methods of testing and evaluation) might well have changed meanwhile. It is necessary, therefore, to define immediate objectives, which are at least not distant in time from the activities of the school.

The formulation and definition of ultimate educational objectives have generally been considered to be the province of educational philosophy. These ultimate objectives must then be analyzed and interpreted in order to formulate the derived immediate objectives of the educational program. Given the derived immediate objectives, it is the function of the curriculum experts to formulate educational programs and procedures which appear likely to achieve them, and it is the function of measurement experts to devise methods for determining the degrees to which they are achieved by various educational programs and by students having all sorts of different patterns of background and ability.

The preceding paragraph suggests by omission what is often the weakest link in the whole process. It does not discuss the procedure by which the immediate objectives are to be derived from the ultimate objectives. It does not name a group whose primary concern is precisely this problem. As a matter of fact, no such group exists at the present time. The problem is everybody's problem, but nobody's particular problem. Yet in the end it is one of the most important of all major educational problems. When an educational program is shown to accomplish its immediate objectives, as it often is, its validity is not thereby established. The ultimate validity of the program is still no better than the validities of the immediate objectives. When evaluations of the program and of the accomplishments of the students are based on immediate objectives, as in practice they must be in most cases, the criteria are *intermediate criteria.* The relevance of the intermediate criteria (the immediate objectives) to the ultimate criteria (the ultimate objectives) must still be demonstrated, either logically or empirically.

All too often the immediate objectives are derived by backward reason-

ing from traditional elements of the curriculum: subjects are justified by showing that the immediate objectives implied by the educational procedures seem to be related to one or more of the ultimate objectives. The immediate objectives are arbitrary, rather than derived. If good tests are built to measure the achievement by students of arbitrary immediate objectives defined in this manner, they help to freeze educational procedures in the traditional molds. If the arbitrary immediate objectives lack ultimate relevance, such tests retard educational progress instead of stimulating it. Test constructors therefore have an ethical obligation to examine and help in formulating the immediate objectives whose accomplishment their tests are to measure.

The derivation of immediate objectives from ultimate objectives is not an easy task, but it is a task of major importance. Partly it must be done, at any given time, by judgment and agreement based on general educational experience. More and more, however, it is becoming one of the main tasks of educational research. Since test development is itself based on research procedures, the measurement experts are as a group at least as well trained in the methodology of educational research as are the members of any other group, and it is quite possible that they are as a whole the best trained group of all in this respect. Hence, they must accept considerable responsibility in formulating the immediate objectives of education, and still more in defining these objectives in terms of observable acts to be performed upon specified materials under clear-cut conditions.

"Ability to do the kinds of arithmetic one is likely to have to do in later life" is one such immediate objective which is much more easily derived from ultimate objectives than are most others. The restatement of this objective in specific terms is not a matter of speculation, judgment, and agreement; it is a matter of research. Some aspects of the research will employ tests, for example, in finding out what arithmetical processes the average adult can handle and what ones he cannot. If it is discovered that most adults no longer use short division, but employ the standard long-division process with one-digit divisors as well as with divisors having more than one digit, it may be advisable to induce teachers to throw out short division, as they have long since thrown out cube root and "casting out nines."

Reading tests and spelling tests are already built routinely on the basis of word counts, which form the basis also for the contents of readers and spellers. Attempts to base instruction in history and geography on fact counts have been less fortunate. The facts do not "stay put" long enough, and facts per se are likely to be overemphasized at the expense of the more

important interpretative ideas and generalizations. We should try to get at these latter directly, by analyzing as best we can the *de facto* uses to which people put educational information, or the uses to which they ought to put such information. We might also find out by research which, among a great many proposed general concepts, are the ones actually possessed by considerable numbers of people who are considered educated by their fellows.

Direct objectives and transfer objectives

One of the major functions of achievement testing is to assist in the evaluation of the educational program. Another is the estimation of the nearness of students to immediate educational goals. Neither of these functions can be performed unless the educational goals are so specified as to imply unambiguously the materials and the actions relative to such materials by means of which the students' nearness to the goals can be appraised. Even the specific statement of educational goals is not enough to permit evaluation of the program. The goals must be given relative weights. The Latin teacher may state that improvement in the effectiveness of use of the English language is one of the goals of instruction in Latin. So far, so good. Both he and the test constructor agree that effective use of the English language is one of the major general objectives of the educational program. "But now," the test constructor asks, "how important is this goal in comparison with the other goals of instruction in Latin?" The Latin teacher may look a bit surprised as he replies, "Just what do you mean? The importance of effective use of English is almost unlimited. So are some of the other aims of Latin instruction. How am I to go about the process of assigning relative weights to the importance of goals whose actual importances are almost absolute?" The reply might be, "My job is to construct a test of achievement in Latin. I propose to include questions covering every area of knowledge and skill and ability covered by the proper objectives of instruction in Latin. What I really want to know is this: if my test contains 150 items, how many of them should be items testing effective use of English?" If the Latin teacher objects at this point, and suggests that improvement in English is a secondary objective rather than a direct objective, and should not be tested at all in a Latin test, the test constructor might not object too strenuously. But he should then point out that improvement in English is now, in fact, a transfer objective of Latin. In the effectiveness of Latin test he will include only those items which test the direct objectives of instruction in Latin. In a separate effectiveness of English test he will include the items based on the direct objectives of instruction in English usage. Then he or someone

else should set up a transfer of training experiment, to find out the extent to which instruction in Latin actually does improve effectiveness of use of the English language. The relative improvement produced by studying English plus Latin, as compared to the improvement produced by studying English only, must then be accepted by the Latin teacher as the *de facto* importance of this objective of instruction in Latin.

If the geometry teacher says that improvement in reasoning ability is one of the major objectives of instruction in his subject, we shall inquire what types of reasoning and over what range of materials. Then we shall insist that he either allocate items covering reasoning in these areas to the geometry examination, or agree to accept the results of transfer of training experiments as an evaluation of the contribution of his instruction to this goal. If he says that his subject is essential to the study of trigonometry and analytic geometry, we may first inquire how many of his students actually do later study these subjects. If the answer is that a large percentage of them do so, we shall ask him to what extent he desires that we evaluate his work by consulting the later records of the work done by his students in analytic geometry and trigonometry. We may also suggest a study in which groups matched on other variables, but one group having studied geometry and the other not, be compared as to later achievement in trigonometry and analytic geometry.

In the case of still other subjects we shall have to insist on the direct inclusion of test items covering the stated secondary objectives. This will happen whenever there is no course or set of courses whose direct objectives are identical with the stated secondary objectives. If the chemistry teacher says he is teaching scientific attitudes, we shall probably insist that he test the scientific attitudes of his students, however crudely. If the art teacher says he is developing artistic appreciation, we shall attempt to test artistic appreciation, and judge the effectiveness of his program and the progress of his students in part thereby. And if the principal of a certain high school believes that the development of general eye-hand coordination is important, and sets up a required course in billiards to develop it, we shall offer no objection so long as he permits us to evaluate its effectiveness by testing eye-hand coordination in a variety of situations rather than merely by testing the students' proficiency at billiards!

The aims of education and the objectives of the school

It is a truism implicit in the phrase "we live and learn" that all human activities result in learning; hence all social institutions have educational objectives. The school is only one such institution, and since its authority is limited, its responsibility must be limited in like degree. Furthermore,

responsibility has no meaning in the absence of some method for determining whether it has been discharged effectively or ineffectively. If the school wants to accept some responsibility, but lacks either the necessary authority to discharge it or the ability to evaluate, however crudely, the results of its proposed efforts, we must conclude that it is unable to accept this responsibility.

Some of the aims of education can and should be stated in terms of acts to be performed by students while actually in school. Such, in general, are the aims associated with what are termed tool subjects. The tools or skills developed are useful and important to children as well as to adults. In these areas it is usually possible with sufficient effort to obtain intrinsically relevant criterion samples, and this should be done. When less direct and more convenient testing procedures can be devised, they should be validated against these intrinsically relevant criteria.

In other areas the desired acts are acts to be performed by students of school age, but these acts—or most of them—occur outside the school itself. Many of the acts which are evaluated in ethical and moral terms come in this category. The school has some responsibility in this area, but it does not have major responsibility because it does not have major authority. Personality and character development is the responsibility of all social institutions, of which the school is only one. Ultimate responsibility for children's moral conduct rests with the parents up to a certain point, and thereafter with the courts. The school has three main responsibilities: (1) to teach the verbal concomitants of the proper acts, so that the child will not perform wrong acts in the belief that they are generally considered right; (2) to see that the child performs the right (and in general the best) kinds of acts while in school; (3) to perform its primary functions (including the two above) in such a manner as to avoid producing conditions in the child which may themselves lead to wrong acts outside of school. But to recognize these latter when and if they occur, it will be necessary for the school to obtain reports of systematic observations of the out-of-school behavior of the children. If it cannot do this, it will have great difficulty in meeting the third responsibility.

Much muddled thinking about aims of this type will be avoided as soon as it is realized that character and personality, insofar as anything can be done about them, consist of acts, and in particular of those acts which are evaluated in ethical and moral and social terms. Such acts can and should be observed and scored. If the school accepts responsibility in a certain subarea, the child's behavior in that subarea should be observed and scored. If the performance is poor, it should be so reported—perhaps as "failure in such-and-such subarea of personality or character development." If too

many children fail, in this as in any other field, the school itself has failed, either in effective teaching or in adapting its teaching methods and materials to the pre-existing backgrounds of the students. If the school is unable or unwilling to observe and report the actions of its students in such a subarea so that the community can in turn evaluate the school's own work, then it is in fact unable or unwilling in like degree to assume the corresponding responsibility, regardless of what may appear in the "aims" section of its course of study.

It is true, of course, that we can use paper-and-pencil tests to measure directly only the verbal concomitants of ethical and moral and social behavior: namely, "knowledge about" such behavior. If teachers are unable or unwilling to use other observational and scoring techniques, the school is in fact limiting its responsibility to this area. Actually, the use of systematic anecdotal records of acts observed, of behavior check lists, and of rating scales will probably make it possible to rate observed acts of ethical and moral and social behavior with some useful validity, if teachers really want to rate such acts. If they do not, they should stop talking about personality and character and citizenship as functioning aims of school education.

Learning may be the direct result of direct teaching of the acts to be learned. It may also be a concomitant of the methods used and of the attitudes exhibited in teaching other acts. If concomitant learnings are considered important, the outcomes of such learnings must be observed and scored. If they are not observed and scored, they are by that fact labeled unimportant. Concomitant learnings may be important intended outcomes of the educational process, or the term may be an empty phrase so far as the actual objectives of instruction are concerned.

Levels of evaluation

In the previous section we have considered educational aims which can be stated in terms of acts actually performed by school children, either in or out of school. But in still other areas, the desired acts are acts to be performed long after the students have finished their formal schooling. The aims of content subjects and of appreciation subjects fall generally in this class. Information is not taught for its own sake (that is, for verbal reproduction of the items), but for the sake of the uses to which it will be put later. It was pointed out in chapter 5 that alternative sets of items of information may be used in building up those general patterns of comprehension that are termed "background." The main features of the patterns often remain after the facts as such are forgotten. At the level of instructional-unit evaluation we may test for knowledge of the facts, on the

assumption that if they are not retained for at least a few days or weeks, the interpretative concepts and organizational syntheses will fail to develop, or that the ones which do develop will be distorted. Even this much fact testing (in content subjects) has been questioned. If we are right in assuming that, within limits, one set of facts is as good as another for the purpose of developing general understandings, we do not have to reach full agreement on just what facts are to be taught, and each teacher's fact tests will be based directly on the facts that *were* taught.

At the level of course evaluation, and quite possibly at the level of unit evaluation also, we should test directly for understandings, interpretative ideas, and organizational concepts. At this point it becomes necessary to define and agree upon the particular understandings, interpretations, and organizational concepts which constitute the course objectives. Once these agreements have been reached, we can test for the understandings, interpretations, and organizational concepts directly. There is no distinction at this level between a test and a criterion sample, but we must be careful to note that the criterion itself is an intermediate criterion, since the objective is an immediate objective. We can say of the student who makes satisfactory scores on our tests that he has *at this time* certain elements of background. The ultimate effectiveness of the teaching and learning, however, is to be judged in terms of his later actions.

The organization of background learning into subjects is more or less arbitrary. Many if not most of the really important organizational patterns, and many interpretative ideas and general understandings as well, cut across courses and even across subject-matter fields. A person cannot possess really useful background in history without possessing some background in geography, economics, political theory, and probably anthropology, psychology, and sociology as well. If his understandings in these areas are compartmentalized, his effective background in each of them will be less than if the same understandings are interwoven. Good teaching of any content subject or appreciation subject consists in encouraging and helping students to perform the acts by which such interwoven understandings are built up, as well as in acquiring the understandings which are the direct concern of the particular course. All this is, of course, an old story to any competent teacher or educator.

If the development of broad understandings and organizational concepts which transcend course boundaries is a proper function of the school, it is up to the school to define these broad understandings and concepts, and more particularly to define the acts which persons who have them are likely to be able to perform, and which persons who do not have them are not so likely to be able to perform. One class of such acts is the solution

of problems which demands the integration of knowledge from several areas. Another is the ability to read passages and interpret them in the light of knowledge of several fields. At this point we are getting fairly close to certain ultimate objectives—ability to assimilate new ideas in terms of an adequate background of old ideas, ability to apply a wide range of understandings and concepts to new problems as they arise, and the like.

From these considerations it follows that there is a third level of evaluation—evaluation of the results of the whole educational program. We cannot escape this conclusion and still insist that there is a "liberal education aim" which transcends the sum of the specific aims of the separate courses. We must insist that the *de facto* aims of an educational program, and of every part thereof, consist of those acts on the basis of which the students and the program are in fact evaluated. If any stated aim is not analyzed into specific actions, and those actions observed and scored and reported, the statement is no more than empty verbiage.

The typical instrument for testing at this third level is the comprehensive examination, which tries to assess the general level of educational development of the student. Such tests have been described in chapter 5. Even at this level the test must usually be validated against intermediate criteria whose relevance to the ultimate aims of the educational program can only be estimated, but which can be improved by more careful definitions of the immediate objectives of a liberal education.

Since the aims of liberal education refer to the future, the ultimate criterion acts are to be predicted, rather than observed while the students are still in school. As far as possible, therefore, we should observe and score random samples of the actions of adults who are graduates of various types of schools. We should score the actions as "informed" actions or "uninformed" actions, or as actions reflecting various degrees of "informedness." The performance of informed acts may be said to characterize an educated person; the performance of uninformed acts, an uneducated person. Eventually we should correlate these ultimate criterion scores with the scores made on educational achievement tests by the same people when they were in school. But if educational achievement tests are well designed and well constructed, so that they actually test directly, or indicate indirectly but with adequate validity the acts which are immediate aims of the liberal education program, the scores they yield *will* be valid with reference to the intermediate criteria. Now, however, we propose to use them as predictors of the ultimate criterion scores. In determining the predictive powers of the test, we are determining the validities of the corresponding parts of the educational program. When we are able to do this

effectively, we will have established the essential foundation for a science of education.

Methods of estimating relevance

It has been noted previously that validity has two aspects: relevance and reliability. Since the reliabilities of most educational achievement tests can be determined fairly readily if their formal structures meet a few fairly straightforward standards, the main problems of validity are the problems of relevance. Whenever it is possible to obtain a set of actual criterion scores along with a set of test scores of the same individuals, we estimate the predictive power of the test directly from the test-criterion correlation, and we can estimate its validity directly also if we are able to separate the criterion scores into two parallel sets. The estimation of relevance is a secondary operation in this case, calling for two parallel sets of test scores in addition. In these cases (involving actual criterion scores) we may say that the test possesses empirical validity to the degree indicated by the statistical index or coefficient. We seldom in practice compute the index or coefficient of relevance of a finished test, though the formulas for doing so are given in the following section of this chapter.

Occasionally, however, we may wish to study the probable value of some new type of test without expending the effort necessary to construct one long enough to possess high reliability. In any such instance we may construct two representative and parallel but short forms of the new test, apply them to a fairly large group for whom two parallel sets of criterion scores are available, and compute the index or coefficient of relevance. If this index or coefficient is high, we may conclude that the new type of test is promising, and go ahead with the development of forms long enough to possess the desired degree of reliability for service use. It should be noted, however, that the less reliable the test (and the criterion scores also), the larger the tryout group must be in order to determine its relevance accurately. The reasons behind this statement, and the exact definitions of "parallel," "index," and "coefficient" are given in the following section on statistical problems.

In a great many situations, the criterion acts are not directly observable. This is true generally of acts of comprehension. The ultimate criterion of comprehension is how it is applied to the solution of problems of all sorts throughout the life span of the individual, but there is no intermediate criterion other than a comprehension test itself. The relevance of such a test must be estimated by comparing the acts evoked by the test with the acts specified by the statement of the immediate objective.

If the test contents, the test operations, and the test situation are essentially those specified by a set of scientifically worked-out derived immediate objectives, the test may be said to possess logical relevance. The test is logically relevant to the derived immediate criteria, and the latter are logically relevant to the ultimate criteria.

If the test contents, operations, and situation are essentially those specified by a set of arbitrary immediate objectives, which are in turn specified in terms of the things actually taught in a unit or course or subject, the test may be said to possess curricular relevance. It is logically relevant to the *de facto* immediate objectives, but the relevance of the latter to the ultimate objectives of education has not been investigated by the test's author.

If the test's contents are essentially those specified by a set of immediate objectives (either derived or arbitrary), but it is not clear that both the test operations and the test situation are those specified by the objectives, the test may be said to possess only formal relevance, which is always to be viewed with considerable distrust. Formal relevance is often mistaken for curricular relevance or even logical relevance by those who are unable or unwilling to make rigorous logical and psychological analyses of the operations and situations implied by general statements of the immediate objectives, and by those who go no further than the analysis of textbooks and assigned readings in the preparation of their test outlines and specifications.

In the case of aptitude, ability, interest, and personality tests we do not usually start with a measurement objective stated in terms of rules specifying the acts belonging to an appropriate criterion series. The items of such a test are devised in accordance with some arbitrary set of standards suggested by psychological theory, and the item analyses are designed usually to make the final test a trait test if this is possible. When such a test has been included in a number of factor analyses, along with a variety of other tests, the results provide us with a more or less accurate description of its factorial relevance to these other tests. The concept of factorial relevance can be extended to batteries which include both tests and criterion scores, but so few such studies have been made as yet that we have little useful information concerning the factorial relevances of tests to criteria or of different criteria to one another.

We should take care to note that the concept of relevance, and that of validity also, applies to the acts evoked by the test materials in the test situation, and not to the test materials (for example, questions) as such. The test materials are in fact only a part—though usually the crucial part—of the total test situation. It is the examinees' responses to

the elements of the test situation, chief among which may be the test questions, and not the situation per se nor any part of it, which possess the properties of intrinsic relevance, validity, and reliability. The final estimates of these properties, and of the qualities of the examinees' test performances, are of course influenced further by any imperfections in the methods by which the responses are recorded, observed, evaluated, and scored.

Logical relevance

An ultimate criterion possesses perfect logical relevance only because the general concept or ultimate educational aim underlying it is defined by a set of rules for the inclusion or exclusion of acts from the criterion series and a set of rules for scoring the acts in the series, and because all the acts in the series can be observed and scored without systematic biases which are common to several observers or scorers. Substitute "whenever" for "because" in the above sentence, and it becomes apparent at once that most ultimate educational aims do not lead to the development of ultimate criteria. The aim, in order to possess unambiguous meaning, should *always* be defined by a set of rules specifying the acts with which it is concerned, but teachers and educators are properly concerned with many series of acts (for example, comprehension again) which are not themselves objectively observable and scorable. In such cases it is often possible to specify subseries of acts, such as comprehension of historical movements and trends in a certain country or group of countries and over a certain period. It is often possible also to specify the quality of comprehension desired in terms of problems which should be solved correctly by persons who possess the desired degrees of comprehension. It is often possible to set up exactly these problems in the form of a test. Such a test, if the specifications are rigorous and exhaustive, possesses logical relevance to that portion of the ultimate objective which is specified by the subseries.

Logical relevance is not necessarily a second choice to empirical relevance in every instance. Consider the construction of a test of job knowledge. This obviously is a part, but only a part, of job proficiency. The series of acts defining job knowledge is not even a subseries of the series defining job proficiency. Job knowledge, like habit, is inferred from interrelations among subseries of the acts constituting job proficiency. In some jobs the interrelations indicate that the knowledge factor is a very large one. In practice it is usually impossible to get rid of all bias in the observation and scoring of the criterion acts, and it is usually impossible also to attain complete agreement, even among one given set

of observers and raters, as to the acts which do and do not fall within the criterion series, and as to the relative importances of those which do. Hence, in practice we seldom deal with a genuine ultimate criterion sample. We deal instead with a working criterion sample which is *judged* to be a sufficient practical approximation to an ultimate criterion sample. In constructing the job knowledge test we have our choice between two types of judgment. If we validate the items statistically, we accept the judgment that the working criterion is adequate, along with all the specific judgments that lead to the criterion scores. We may, alternatively, ask those who know the job to *list* the concepts which constitute job knowledge, and to rate the relative importances of these concepts. Then when the preliminary test is finshed, we may ask them to examine its items. If they agree fairly well that the test items will evoke acts of job knowledge, and that these acts will constitute a representative sample of all such acts, we may be inclined to accept these judgments. There is no evidence that judgments about a set of observed acts of individuals are in general either more or less relevant than are judgments about a set of concepts or judgments about a set of items. Which set of judgments we are to prefer in any given case is itself a matter of judgment, at least until research provides some dependable evidence bearing on this problem.

Often we are called upon to make action judgments on the basis of the best available tests in situations wherein we do not know what tests are the best available, nor the validities of any tests for the purpose at hand. Such situations are the rule, rather than the exception, in educational and vocational guidance, and most of the tests which are used in guidance are intelligence tests, aptitude tests, interest tests, personality tests, and the like, rather than educational achievement tests. How shall we select the best available tests for use in a new and in some degree unique guidance situation? How shall we estimate, however crudely, the validities of such tests? And finally, what can be done to improve this situation, pending the millennium when we shall have a test of known validity for every important purpose?

We cannot well refuse the job, except in those instances where we believe the action judgments will be better without the aid of tests than with them. Johnny is going to enroll in either the technical curriculum or the commercial curriculum. Mary is going to look for either a clerical job or a sales job. Bill is going to major in either education or business administration. Dean Jones is going to hire one of the four girls who have applied for a position as his secretary. We believe that the judgments on which such actions are to be taken will be improved if

they are based in part on test data, but we do not know the precise validities of any tests for these particular purposes. What we do know is that certain tests have been demonstrated to be reasonably valid in somewhat similar situations for somewhat similar purposes. We are not compelled to defend the proposition that the tests recommended will very probably have *high* validities, but we are required to defend, usually without any direct evidence, the proposition that the inclusion of such tests among the items of evidence used will probably improve the quality of the action judgments.

The more the test user knows about the available tests and how well they have worked in a variety of other situations, the better his judgments on these matters are likely to be. The authors and publishers of intelligence, aptitude, interest, and personality tests can assist test users very materially in two ways: (1) by making factor analyses and reporting the factorial compositions of their tests; (2) by searching the literature and reporting all known validation studies in which their tests have been used. These reports should include particularly the studies in which their tests were found to possess low validities, as well as the studies in which the validities of their tests were found to be high. Considerations of factorial composition are of particular value in assembling a reasonably small battery from a large number of tests all of which are known to have been fairly valid in situations similar in various respects to the one at issue.

Judgments concerning the logical relevance of any test for any purpose, with perhaps the exception of tests of knowledge and comprehension, each built to measure one specified area of such knowledge and comprehension, are beset with a variety of dangers. Experience indicates, however, that some types of judgments are likely to be worse than others. Identical elements and related behavior tests can often be assumed quite reasonably to possess fair logical relevance on the basis of evidence concerning their curricular or even formal relevances. On the other hand, the assumption that a verbalized behavior test will evoke relevant behavior because the described situations are representative of typical real situations is quite dangerous. This is one case in which formal relevance may be a very poor indicator of empirical relevance. Tests of the verbalized behavior type have been used to measure supervisory ability, teaching ability, and various aspects of interest and personality. Some of them have been found to work fairly well, and others have not, Neither verbalized behavior nor knowledge tests can safely be assumed to be valid indicators of what people will actually do in response to ethical or moral or social situations, though they *can* in many cases be assumed to possess logical

relevance for measuring what such people think they ought to do. As indicators of probable action, however, the relevances and validities of such tests must be determined empirically before they can be used with any real confidence.

An indirect test may have demonstrated its validity in a variety of situations which more or less "surround" the one at issue. In such cases the assumption that it will be valid in this situation is fairly safe. A study of the literature on the validity of the Bennett Mechanical Comprehension Test led to several fairly clear generalizations. This test almost always discriminated with some significance between good and poor workers in highly skilled mechanical occupations—the ones which require mechanical know-how. It discriminated between good and poor students in high-level shop courses and related science courses. It did not discriminate uniformly between good and poor workers in semiskilled occupations demanding mainly certain manual skills, nor between good and poor students in low-level shop courses designed to develop such skills. With this information we can feel safe in using the results from the test fairly generally in counseling students who are considering careers as electricians, tool and die makers, and the like, but much less generally in counseling those who are considering jobs as radio assemblers, machine operators, and in similar occupations.

The scores on verbal intelligence tests discriminate between good and poor clerks in a wide variety of clerical occupations. The scores on arithmetic tests and on checking tests show quite high validities in some cases and quite low validities in others. In a number of studies in which the criteria consisted of over-all ratings of secretaries by their chiefs, group intelligence tests were more discriminating than were tests of stenographic and typing proficiency. With these points in mind we can probably provide Dean Jones some effective help in employing a secretary.

In the case of many aptitude, ability, interest, and personality tests, there is considerable evidence that tests which would appear to many users to be measuring closely allied functions may or may not actually be doing so. At least in the present state of knowledge, judgments concerning the functions actually measured by such tests are likely to be quite untrustworthy. On the other hand, tests whose factorial compositions have actually been demonstrated to be similar usually show fairly similar correlations with complex criterion performances. The importance of factor analysis studies to determine the functions actually measured by these tests is therefore considerable. Lacking such studies, it is quite hazardous to use a test in a new situation unless it has been shown to be valid in a

variety of similar situations. This puts a premium on the continued use of old tests whose validities in many situations are a matter of record, rather than newer tests which may well be much superior to the old ones but have not been used widely enough to demonstrate this fact with sufficient certainty.

In using a new indirect test in a new situation, we are compounding the dangers inherent in the assumption of relevance. First we must assume that other tests would be relevant in the new situation. Then we must assume that the new test will be relevant in this situation because it is apparently similar to these other tests. Widespread use of the first type of assumption is inevitable. There are too many important criterion situations and performances to permit direct determination of the validities of proposed tests in all or any major fraction of them. On the other hand, the results of factor analysis to date indicate that a comparatively limited number of indirect aptitude, ability, interest, and personality tests could do most of the work now done with a fairly large number of such tests, but a number which is still small in comparison to the number of important criteria.

The assumption of relevance is even more precarious when criterion performances are judged similar on the basis of job qualification analyses rather than on the basis of observed identity or similarity of the actual operations performed. As evidence that we are not knocking down a straw man in making this statement, we may quote from a testing handbook circulated rather widely in one of the military services during World War II:

> Generally speaking, the validity of the test is best determined by using common sense in discovering that the test measures component abilities which exist both in the test situation and on the job. This common-sense approach to the problem can be strengthened greatly by basing the estimate of the components of the job on a systematic observation or job analysis.

If common sense cannot provide a useful description of the factorial composition of a test, it seems very unlikely that it can provide any useful description of the factorial composition of a job, which is usually much more complex. At present we have only the beginnings of a substantial literature on the factorial compositions of indirect tests. There is virtually no literature as yet on the factorial compositions of jobs or of learning tasks, and the basic problems in this area of research are much more complex than are the corresponding problems relating to tests. Similarity between job and job, or test and test, or test and job can be inferred reasonably when the materials dealt with, the operations performed, and the situations in which they are performed can be observed

to be in some considerable part identical or very similar; and under no other conditions.

Factorial relevance

Indirect tests are of great importance because of their convenience and simplicity in application, and because one such test is often useful for a variety of purposes. This is particularly true of tests used to predict future criterion events. In prediction we seek to ascertain the present characteristics of persons who later "succeed" in some job or performance, as compared with the present characteristics of those who later "fail." We have good reason to believe that the crucial characteristics are less numerous than the number of different types of criterion events and performances to be predicted. Some characteristics are important indicators of successes of many types. A pure test of such a characteristic is of great value, in terms of the limitations of time and expense incident to practical testing programs, even though a test measuring this characteristic in terms of the particular materials and operations of each criterion might be somewhat better in every individual instance. The superiority might still be insufficient to justify the expense required to prepare all the separate tests; and if they were prepared, the potential examinees might not have time to take so many of them. Hence, in practice it is valuable to know what intelligence, aptitude, ability, skill, interest, attitude, and personality tests "really" measure.

"Reality" here, as in all other matters, is a relative affair. As more and more different kinds of tests are produced and their intercorrelations are determined, we may reasonably hope that the methods of factor analysis will provide increasingly more meaningful and useful answers to our questions concerning what the various types of psychological tests really measure. More primitive information of the same type about any such test is provided by simply reporting its correlations with a variety of other tests. The authors of psychological tests would be helping the users of such tests considerably if they would report as many of these correlations as possible. They would provide still more assistance if they would subject their tests to as many factor analyses as practical considerations permit, using different combinations of reference tests in the various batteries, and samples from several different populations.

In interpreting and using the results of factor analysis, we should guard against reading more into the results than the data warrant. Factor analysis is simply a method for interpreting the meaning of each test in a battery with reference to all the others. The concepts of abilities and traits as entities, or as essential qualities or characteristics of individuals,

are more often misleading than helpful, since they tend to encourage too much generalization and extrapolation from the actual test results. At the present state of our knowledge concerning the organization of human reaction systems, the interpretation of performances on indirect tests is safe only so long as it remains strictly empirical. A "factor" exists in a table of intercorrelations rather than in a person. The table of intercorrelations applies to a particular group at the time they are tested. Factorial patterns vary considerably among radically different groups, including groups (particularly of children) of radically different ages. There is evidence that such patterns do not vary greatly among groups which are only slightly different, but this evidence is far from conclusive. Hence, only if the results of *several* factor analyses agree in indicating that two tests have *very similar* factorial compositions, and one of them has a known correlation with a criterion in a group which is similar in general to the groups used in making the factor analyses, only then can we conclude that the other is likely to have a fairly similar correlation with that criterion in any group of approximately the same type. But if, say, an arithmetic reasoning test measures only factors which are also measured by a vocabulary test, an arithmetic computation test, and a symbolic reasoning test, and has a small "specific factor," the addition of this arithmetic reasoning test to a battery already containing the other three will not markedly improve that battery's correlation with any criterion in any group similar to the one used in making the factor analysis.

Curricular relevance

This concept, which has sometimes been termed "curricular validity" or "content validity," applies primarily to educational achievement tests. It is a special case of logical relevance, and is based on the assumption that in certain cases there is no need to inquire into the relevance of an immediate objective to the ultimate objectives of education, or that the relevance of the immediate objective has already been demonstrated.

An ordinary subject-matter test has usually been considered to possess curricular relevance to the extent that it tests the students' knowledge and effective grasp of those facts, principles, relations, patterns, and generalizations which are the *de facto* immediate objectives of instruction. The usual evidence of curricular relevance is a tabulation showing that the test content actually parallels and covers the course content; that the test operations (reproduction of facts, application of principles, perception of significant relations, etc.) are those specified in the statement of the course objectives; and that the test situation is not such as to bias the

responses. If a teacher presents the 100 addition facts to her first- or second-grade pupils as a test, someone may ask, "How do you know that this test measures exactly what it ought to measure?" The teacher may rightly smile as she answers that she has just been teaching the children to add pairs of one-digit numbers, and is now asking them to add *all* the pairs of numbers she has just been teaching them to add, and the test is therefore obviously relevant. By this she means that its curricular relevance is not open to any serious question on the basis of the fact that the additions were performed in school rather than somewhere else. The materials and operations are, of course, identical with the ones taught. To be sure, the test may need to be repeated several times, with the orders of the number-pairs varied, before she can be sure which errors are accidental and which represent actual lack of knowledge. But if the test materials, situations, and performances themselves represent all of the pertinent specific objectives of instruction, no further evidence of its curricular relevance is required.

When the objectives of a unit or course or subject can be stated explicitly in terms of knowledge, skill, information, and the like, it is usually possible to construct tests measuring the achievement of these objectives by the students, and such tests possess high curricular relevance. Sometimes the items will be paper-and-pencil items, and sometimes not. If the school's contribution to good citizenship consists in the development of pupil insight into civic and political problems and the accompanying ethical problems, the corresponding citizenship test will possess curricular relevance if it tests the pupils' insight into the problems that were taught. If it tests their insight into a representative sample of the problems which, according to some valid standard, *should* have been taught, it possesses logical relevance.

The curricular relevance of the 100-item addition test mentioned previously is perfect. This is, of course, an exceptional case, though it can be extended to cover almost the whole of arithmetic computation. Curricular relevance, however, does not necessarily imply complete coverage of the field. A random sample or a properly controlled stratified sample of the instructional materials may be quite sufficient. Consider the case of spelling. A spelling test will possess perfect curricular relevance if it is based on the words specified by a course of study, or on some other specified list, and on operations at least similar or equivalent to those used by the pupils in spelling the words when they need to use them. If the test consists of a random sample of all the words in the complete list, or better yet a proportional sample of all the words at each level of difficulty, it may be reasonably claimed to possess almost perfect

curricular relevance. Actually, however, a spelling test would seldom be put together in this fashion. Its author would try to improve its reliability and its discriminating power by using only words which are fairly frequently misspelled, and, among these, words whose correct or incorrect spelling can be shown to be correlated with the correct and incorrect spelling of other words. Such a test would not necessarily possess high curricular relevance, but it might be validated against a criterion list of high or almost perfect curricular relevance. It could then be claimed to possess demonstrated curricular validity; that is, curricular relevance plus reliability. It would be more efficient than the test possessing direct curricular relevance: it would be much shorter than a list of the latter type having equal reliability, and yet correlate equally or almost equally with another long test having direct curricular relevance but consisting of a different sample of words from the original large list.

As we move from ordinary unit and course and subject achievement tests to tests covering larger areas of knowledge and behavior, and from these to still more comprehensive tests which attempt to assess the general level of the students' educational development, the value of curricular relevance becomes progressively lower. The only safe standards of relevance for such tests are the standards of logical relevance developed on the basis of research from explicit definitions of the ultimate aims of education. Such tests should always measure genuine desired outcomes of the educational process rather than merely the outcomes to be expected from existing instructional materials and methods.

The logical relevance of even an identical elements test will usually be less than perfect, as has been shown in chapter 5, because it is usually a biased sample rather than a representative sample of the criterion behavior. As we shift our consideration to related behavior, verbalized behavior, and knowledge tests, the gap between logical relevance and curricular relevance becomes progressively greater for all objectives other than those directly concerned with knowledge, and the necessity for careful and astute judgments in setting up immediate objectives becomes correspondingly more important. Test constructors must take care that the immediate objectives they use in developing their test plans and specifications are based in every case upon the best possible analyses of the ultimate objectives and upon resyntheses based on all of the available data supplied by educational research. They should be particularly critical of immediate objectives which are derived by synthesis from existing curricular materials and instructional practices. Only thus can they produce tests whose logical relevances are genuine and adequate.

In practice, tests are often selected and used for some fairly limited

purpose rather than constructed especially to measure the attainment by pupils of the immediate objectives of any particular unit or course or program of instruction. Thus, an algebra teacher may use a test of arithmetic comprehension to find out whether the students know the arithmetical operations that will normally be taken for granted in teaching algebra. Such a potential test user may decide, after examining the items of a test, that it does appear to measure some one or more areas of achievement concerning which he would like to know the relative standings of the members of some group of pupils, or their average standing as compared to other groups. Since curricular relevance, like all other types, is always a matter of relevance for some purpose, it is perfectly reasonable for the test user to propose to use a test to measure what it actually does measure. In a certain sense the test then has perfect curricular relevance, or even perfect relevance (without qualification), because the user is using it to find out precisely what it actually tells him. In such cases its reliability is the only other item of information about it with which he needs to be concerned. In order to assist such potential test users to make informed decisions, the builders of achievement tests should give in their manuals complete statements of the test outlines and specifications, so that the *de facto* nature of the functions measured by their tests can be appraised as accurately as possible.

Formal relevance and "face validity"

The dangers of mistaking formal relevance for curricular relevance or logical relevance have already been pointed out, and need only be recalled briefly at this point. A subject-matter achievement test may cover the outline of a course or of an immediate instructional objective quite thoroughly and with due regard to correct emphasis on the various topics, yet consist almost entirely of questions calling merely for the reproduction of facts and failing to measure the broader aspects of comprehension. A verbalized behavior test may similarly measure knowledge *about* some ethical or moral or social issue, but such knowledge may itself have only limited relevance to actual behavior.

In the areas of aptitude, ability, interest, and personality measurement the same concept is commonly termed "face validity." A test is face-valid if it looks valid—particularly if it looks valid to laymen. Face validity is often important in the public relations aspects of certain types of test programs, but as a validity concept it merely reflects inadequate or superficial analysis. So long as we realize that face validity is not logical relevance, no harm need result from attempts to make tests face-

valid to increase their public acceptability, provided this does not result in weakening their logical or empirical relevances (7).

Validity and criterion reliability

It is usually fairly easy to construct a test whose reliability is high, but much harder (and in most cases impossible) to construct one whose relevance to a criterion approaches perfection. The criterion behavior may be very difficult (but not impossible) to observe and score, in the sense that every observation is troublesome and costly, and that a great deal of time would be required to obtain a reliable sample of behavior for each member of the validation group. In this case we may go to great pains to obtain a representative sample of the behavior in every case, so that the intrinsic relevance of the working criterion scores will be close to perfect. The sample of behavior for each person, however, may consist necessarily of a very few such representative observations, and hence possess low reliability. The raw test-criterion correlation must then be still lower (except by chance), but on correcting it for criterion attenuation we may find that the validity of the test actually exceeds the reliability of the criterion sample scores, despite the fact that its relevance is appreciably less than perfect. The validity of an intrinsically relevant criterion (that is, an ultimate criterion) is its own reliability, but the upper limit of the validity of a test (when it is perfectly relevant to the criterion) is the geometric mean of the criterion reliability and the test reliability. Hence, if the test is much more reliable than the criterion, its validity may exceed the criterion reliability while its relevance remains less than perfect.

Suppose we define school achievement arbitrarily as the sum total of all non-chance influences which produce movement from lower grades to higher grades. By this arbitrary definition, grade placement is the ultimate criterion of school achievement. Consider next the end of each school year, when the child is held back, promoted normally, or double-promoted. A demotion subsequent to a promotion is considered a hold-back, and an extra promotion during the school year is considered equivalent to a double promotion at the beginning of that year. At the end of each school year we give each child a criterion item score of 0 if he is held back, 1 if he is promoted, and 2 if he is double-promoted. Summing his criterion item scores separately for the odd and even school-year-ends, we obtain two sets of criterion sample scores, from which the criterion reliability may be computed. Then we give any of the better general school achievement tests to all the children in grades four to

nine, inclusive, correlate the test scores with the criterion (grade placement), and correct for criterion attenuation to obtain the test validity. The test reliability will greatly exceed the criterion reliability, and in one study in which this procedure was followed, the test validity also exceeded the criterion reliability. Both the odd-year promotion scores and the even-year promotion scores correlated higher with the test scores than they did with each other. The test was a better indicator of school achievement as defined than was the actual grade placement.

Statistical Problems of Validity

From time to time in the preceding discussion we have mentioned various statistical formulas which are used in estimating the several types of validity of test scores. Most of these formulas involve correlations between tests and criterion samples. Correlation formulas, and other statistical formulas as well, are derived on the basis of postulates or assumptions. If the formulas are to be applicable, the experimental designs must be those implied by the assumptions. This point is extremely important; its neglect in the past has led to much research, the results of which cannot be interpreted or summarized properly. Not knowing this, the authors have often interpreted their results improperly. In this section we shall try to set forth the principal formulas used in studying the validities of tests, and attempt to show not only the conditions under which they are applicable, but also some of the conditions under which they are not applicable but have been used erroneously in the past.

The estimation of predictive power

The reliability coefficient was defined in chapter 15 as a "variance ratio": the ratio of the variance (squared standard deviation) of a set of "true" scores to the variance of the corresponding actual scores. The index of reliability was defined as the square root of the reliability coefficient, which means that it could also have been defined as the standard deviation ratio of "true" scores to actual scores. We shall generalize this terminology, defining the validity coefficient, the relevance coefficient, and the prediction coefficient as variance ratios; and the index of validity, the index of relevance, and the index of prediction as standard deviation ratios.

There is a general set of relationships between any variance ratio or standard deviation ratio and the corresponding correlation coefficient. These relationships are most easily demonstrated when both variables are measured directly, as they are when we predict a set of actual criterion

scores on the basis of a set of actual test scores. We define the *prediction coefficient* as the ratio of that part of the criterion variance determined by (and hence predicted by) the test variance to the total criterion variance. From elementary correlation theory we know that

$$\frac{{}_1s_2^{\,2}}{s_2^{\,2}} = 1 - \frac{s_{2.1}^{\,2}}{s_2^{\,2}} = r_{12}^{\,2}, \tag{1}$$

where variable 1 is the test,
variable 2 the criterion,
${}_1s_2^{\,2}$ that part of the criterion variance which is associated with (or determined by) the test variance,
$s_{2.1}^{\,2}$ that part of the criterion variance which is statistically independent of the test variance (that is, uncorrelated with the test variance),
$s_2^{\,2}$ the total criterion variance, and
r_{12} the test-criterion correlation.

The expression ${}_1s_2^{\,2}/s_2^{\,2}$ or $r_{12}^{\,2}$ is ordinarily termed the coefficient of determination; according to our previous definition it is the *coefficient of prediction*. Given a correlation coefficient, we may of course predict the value of either variable from the value of the other, and (1) could just as well have been written,

$$\frac{{}_2s_1^{\,2}}{s_1^{\,2}} = 1 - \frac{s_{1.2}^{\,2}}{s_1^{\,2}} = r_{12}^{\,2}. \tag{1a}$$

In (1a), ${}_2s_1^{\,2}/s_1^{\,2}$ is also the coefficient of determination, and ${}_2s_1^{\,2}/s_1^{\,2} = {}_1s_2^{\,2}/s_2^{\,2}$, since both are equal to $r_{12}^{\,2}$. The coefficient of determination indicates the *proportion* of the variance of each variable which is directly associated with the variance of the other; it does not indicate which determines which, nor in fact anything about the causal nature of the association.

We can also define the *index of prediction* as a *standard deviation ratio* corresponding to the variance ratio which defines the prediction coefficient. Then from (1),

$$\frac{{}_1s_2}{s_2} = \sqrt{1 - \frac{s_{2.1}^{\,2}}{s_2^{\,2}}} = r_{12}. \tag{2}$$

The index of prediction is simply the test-criterion correlation coefficient.

The expression $s_{2.1}^{\,2}$ is termed the variance error of estimate; its square

root, $s_{2.1}$, is the standard error of estimate; $s_{2.1}^2/s_2^2$, which from (1) is equal to $1 - r_{12}^2$, is the coefficient of nondetermination; $s_{2.1}/s_2$, which is equal to $\sqrt{1 - r_{12}^2}$, has been termed the coefficient of alienation by Kelley, though a term more consistent with the others would be the *index of nondetermination.*

From (1a), we can also write,

$$\frac{{}_2s_1}{s_1} = \sqrt{1 - \frac{s_{1.2}^2}{s_1^2}} = r_{12}, \tag{2a}$$

and ${}_2s_1/s_1 = {}_1s_2/s_2 = r_{12}$. From (1), (1a), (2), and (2a), we can draw the following generalizations:

a) Every variance ratio is equal to a second variance ratio defined by interchanging the variables, and to the square of a correlation coefficient.

b) Every standard deviation ratio is equal to a second standard deviation ratio defined by interchanging variables, and to a correlation coefficient.

From these generalizations it is clear that any function (such as the prediction coefficient) defined by a variance ratio can also be defined by the square of a correlation coefficient, and any function (such as the index of prediction) defined by a standard deviation ratio may also be defined by a correlation coefficient.

The estimation of reliability

The coefficients of reliability, validity, and relevance may all be defined in terms of variance ratios which involve the variances of sets of "true" scores. Since the "true" scores cannot be measured directly, it is necessary to estimate their variances from relations among the actual scores. This ordinarily requires two or more sets of actual scores, obtained under certain specific limiting conditions. We shall consider these conditions first in connection with the reliability coefficient, which has been defined as the variance ratio of a set of "true" scores to the actual scores which measure the same function (apart from error). Let x represent a "true" score, and let x_1 and x_{I} represent the corresponding actual scores on two forms of a test, both of which measure x and nothing else except random errors e_1 and e_{I}. Since these errors are random by definition, they must be uncorrelated with each other and with x, and x must be the sole cause of the correlation between x_1 and x_{I}. Experimentally this situation is reversed. We can *compute* only $r_{1\mathrm{I}}$, and it follows from the definitions that x is whatever is common to x_1 and x_{I}, causing them to be correlated;

e_1 is whatever is unique to x_1, causing part of its variability but none of its correlation with x_{I}; and e_{I} is whatever is unique to x_{I}, causing part of *its* variability but none of its correlation with x_1. All of these quantities, x_1, x_{I}, x, e_1, and e_{I}, may be taken as measurements made from their respective means as origins; that is, as "deviation scores." The two sets of scores may be specified as follows:

$$x_1 = c_1(x + e_1); \qquad x_{\mathrm{I}} = c_{\mathrm{I}}(x + e_{\mathrm{I}}). \tag{3}$$

From our definition of errors as uncorrelated with the true scores,

$$s_1^2 = c_1^2 s_x^2 + c_1^2 s_{e_1}^2; \qquad s_{\mathrm{I}}^2 = c_{\mathrm{I}}^2 s_x^2 + c_{\mathrm{I}}^2 s_{e_{\mathrm{I}}}^2. \tag{4}$$

The constants, c_1 and c_{I}, are necessary in order to avoid the assumption that differences between s_1^2 and s_{I}^2 are due directly to differences between $s_{e_1}^2$ and $s_{e_1}^2$. Both the variance of a test and its reliability are in general related directly to its length, but if we omitted c_1 and c_{I} we should conclude from (4) that the form having the greater raw-score variability must be the less reliable.

By (1) and the discussion immediately preceding (3), we ought to define the reliability coefficient of each form of a test as the ratio of that part of the "true" variance determined by the actual variance to the total "true" variance. Letting R_1 and R_{I} be the reliability coefficients of forms 1 and I respectively,

$$R_1 = \frac{{}_1s_x^2}{s_x^2} = r_{1x}^2; \qquad R_{\mathrm{I}} = \frac{{}_{\mathrm{I}}s_x^2}{s_x^2} = r_{\mathrm{I}x}^2. \tag{5}$$

Ordinarily, however, the reliability coefficient has been defined as in chapter 15, as the variance ratio of "true" scores to actual scores. From this definition and (3),

$$R_1 = \frac{c_1^2 s_x^2}{s_1^2} = 1 - \frac{c_1^2 s_{e_1}^2}{s_1^2}; \qquad R_{\mathrm{I}} = \frac{c_{\mathrm{I}}^2 s_x^2}{s_{\mathrm{I}}^2} = 1 - \frac{c_{\mathrm{I}}^2 s_{e_{\mathrm{I}}}^2}{s_{\mathrm{I}}^2}. \tag{6}$$

To prove that these definitions are consistent, we note first that either of two correlated variables may be predicted by the other, and the coefficient of determination is the same in both cases. Then from (1), reversing the substitution which led to (5),

$$\frac{{}_xs_1^2}{s_1^2} = 1 - \frac{s_{1.x}^2}{s_1^2} = r_{1x}^2; \qquad \frac{{}_xs_{\mathrm{I}}^2}{s_{\mathrm{I}}^2} = 1 - \frac{s_{\mathrm{I}.x}^2}{s_{\mathrm{I}}^2} = r_{\mathrm{I}x}^2.$$

But $s_{1.x}^2$ and $s_{\mathrm{I}.x}^2$ are those parts of s_1^2 and s_{I}^2 which vary independently of s_x^2; that is, they are $c_1^2 s_{e_1}^2$ and $c_{\mathrm{I}}^2 s_{e_{\mathrm{I}}}^2$ as defined by (4). Hence,

$$r_{1x}^2 = 1 - \frac{c_1^2 s_{e_1}^2}{s_1^2}; \qquad r_{\mathrm{I}x}^2 = 1 - \frac{c_{\mathrm{I}}^2 s_{e_{\mathrm{I}}}^2}{s_{\mathrm{I}}^2}.$$

Since the second term of each of these equations is also one of the terms of the corresponding equation of (6), it follows that the definitions of R_1 and R_{I} given by (5) and (6) are consistent; they both lead to the relations, $R_1 = r_{1x}^2$, and $R_{\mathrm{I}} = r_{\mathrm{I}x}^2$.

We must consider further the restrictions implied by the definition that x is the same "true" ability in x_1 and x_{I}, even though the two forms be unequal in length and unequal in either the absolute amounts or the relative proportions of their respective errors. We shall suppose that x is made up of two or more uncorrelated underlying abilities. If the real underlying abilities are correlated, they may always be represented mathematically by a hypothetical set of the same number of uncorrelated abilities. Assuming two such abilities, a and b, with weighting constants k_1 and k_{I} for a and m_1 and m_{I} for b, we have from (3),

$$x_1 = c_1(x + e_1) = c_1(k_1 a + m_1 b) + c_1 e_1. \tag{7}$$

$$x_{\mathrm{I}} = c_{\mathrm{I}}(x + e_{\mathrm{I}}) = c_{\mathrm{I}}(k_{\mathrm{I}} a + m_{\mathrm{I}} b) + c_{\mathrm{I}} e_{\mathrm{I}}. \tag{8}$$

Unless $k_1 = k_{\mathrm{I}}$ and $m_1 = m_{\mathrm{I}}$, we have in these equations two inconsistent definitions of x. This should be intuitively apparent, and it can be proved by showing that in the absence of these equalities the correlation between $(k_1 a + m_1 b)$ and $(k_{\mathrm{I}} a + m_{\mathrm{I}} b)$ will not equal unity. The meaning of the required equalities may be clarified by considering the mean and variance of $c_1 x$ and $c_{\mathrm{I}} x$, and the reliabilities R_1 and R_{I}. From the first two terms of (7) and (8), since positive and negative errors of measurement will occur with equal frequency if the errors are random, $M_1 = c_1 M_x$, and $M_2 = c_2 M_x$. Then from the third terms,

$$M_1 = c_1(k_1 M_a + m_1 M_b); \qquad M_{\mathrm{I}} = c_{\mathrm{I}}(k_{\mathrm{I}} M_a + m_{\mathrm{I}} M_b).$$

If $k_1 = k_{\mathrm{I}}$ and $m_1 = m_{\mathrm{I}}$, the means of a in forms 1 and I must be proportional to M_1 and M_{I}, and the means of b in forms 1 and I must also be proportional to M_1 and M_{I}. From (7) and (8), also,

$$c_1^2 s_x^2 = c_1^2(k_1^2 s_a^2 + m_1^2 s_b^2); \qquad c_{\mathrm{I}}^2 s_x^2 = c_{\mathrm{I}}^2(k_{\mathrm{I}}^2 s_a^2 + m_{\mathrm{I}}^2 s_b^2).$$

If $k_1 = k_{\mathrm{I}}$ and $m_1 = m_{\mathrm{I}}$, the variances of a in forms 1 and I must be proportional to $c_1^2 s_x^2$ and $c_{\mathrm{I}}^2 s_x^2$; and the variances of b in forms 1 and I must also be proportional to $c_1^2 s_x^2$ and $c_{\mathrm{I}}^2 s_x^2$. Then from (6),

$$R_1 = c_1^2 \frac{k_1^2 s_a^2 + m_1^2 s_b^2}{s_1^2}.$$

$$R_{\mathrm{I}} = c_{\mathrm{I}}^2 \frac{k_{\mathrm{I}}^2 s_a^2 + m_{\mathrm{I}}^2 s_b^2}{s_{\mathrm{I}}^2}.$$

Expanding these two equations,

$$R_1 = c_1^2\left(\frac{k_1^2 s_a^2}{s_1^2}\right) + c_1^2\left(\frac{m_1^2 s_b^2}{s_1^2}\right).$$

$$R_{\mathrm{I}} = c_{\mathrm{I}}^2\left(\frac{k_{\mathrm{I}}^2 s_a^2}{s_{\mathrm{I}}^2}\right) + c_{\mathrm{I}}^2\left(\frac{m_{\mathrm{I}}^2 s_b^2}{s_{\mathrm{I}}^2}\right).$$

If $k_1 = k_{\mathrm{I}}$ and $m_1 = m_{\mathrm{I}}$, the proportionality of variances implies a corresponding proportionality of reliabilities; that is, the reliability with which a is measured in forms 1 and I must be proportional to the total reliabilities of forms 1 and I, and the reliability with which b is measured in forms 1 and I must also be proportional to the total reliabilities of forms 1 and I.

If the abilities a and b are measured by two independent subsets of items in each of the two forms of the test, the requirement of proportional means indicates that the average difficulty of the a items must be the same in forms 1 and I, and that the average difficulty of the b items must also be the same in forms 1 and I. The requirement of proportional reliabilities indicates that the average interitem correlation of the a items must be the same in forms 1 and I, and that the average interitem correlation of the b items must also be the same in forms 1 and I. The requirement of proportional variances, in addition to the other two requirements, is *almost* sufficient to justify the following summary statement: the joint distribution of the item difficulties and the internal-consistency item discriminations of the a items must have all constants except n, the number of items, the same in form 1 as in form I, and this statement is also true as regards the b items. To clinch this argument, it is necessary to show that all the higher moments and product-moments must also be proportional if $k_1 = k_{\mathrm{I}}$ and $m_1 = m_{\mathrm{I}}$; this can be done by methods essentially similar to the preceding, though more laborious.

We need a name to designate pairs or sets of scores or test forms which conform to these requirements; that is, which measure the same proportional combination of "true" abilities, but not necessarily with equal reliability. We shall call this property "parallelism."

Parallel sets of scores or test forms are sets of scores or test forms which measure the same proportional combination of "true" abilities, but not necessarily with equal reliability.

The preceding somewhat technical discussion has centered around two points: (1) the proof that two alternative definitions of the reliability coefficient are consistent, and (2) an analysis of the meaning of the phrase "the same proportional combination of 'true' abilities" as used

in the definition of parallel sets of scores or test forms. The remaining formulas to be considered are all well known, and will be presented without derivations.

We have noted previously that defining formulas which include variances or standard deviations or correlations of "true" scores do not lead to direct computations, since "true" scores cannot be measured directly. Hence, we must estimate their values from the variances, standard deviations, and correlations of the raw scores.

If R_1 and R_I are the reliability coefficients of parallel forms of a test,

$$R_1 R_I = r_{1x}^2 r_{Ix}^2 = r_{1I}^2. \tag{9}$$

Similarly if r_1 and r_I are the indices of reliability,

$$r_1 r_I = \sqrt{R_1 R_I} = r_{1x} r_{Ix} = r_{1I}. \tag{10}$$

We next define "equivalence" as follows:

Equivalent sets of scores or test forms are parallel sets or forms which are equally reliable.

Then if x_1 and x_I are equivalent forms,

$$R_1 = R_I = r_{1x}^2 = r_{Ix}^2 = r_{1I}. \tag{11}$$

$$r_1 = r_I = r_{1x} = r_{Ix} = \sqrt{r_{1I}}. \tag{12}$$

These are the usual formulas for the reliability coefficient and the index of reliability.

The estimation of validity

The validity coefficient may be defined as the ratio of that part of the "true" criterion variance determined by the test variance to the total "true" criterion variance. In order to evaluate this ratio we require *two parallel,* but not necessarily *equivalent,* sets of criterion scores. Let y be a "true" criterion score, and let y_2 and y_{II} be the two corresponding actual scores, x_1 being a score on the test whose validity is to be estimated. The validity coefficient is

$$\frac{{}_1s_y^{\,2}}{s_y^{\,2}} = r_{1y}^2 = \frac{r_{12} r_{1II}}{r_{2II}}. \tag{13}$$

The index of validity is

$$\frac{{}_1s_y}{s_y} = r_{1y} = \sqrt{\frac{r_{12} r_{1II}}{r_{2II}}}. \tag{14}$$

The estimation of relevance

The relevance coefficient may be defined as the ratio of that part of the "true" criterion variance determined by the "true" test variance to the total "true" criterion variance. In the previously defined terminology, the relevance coefficient is

$$\frac{{}_{x}s_{y}^{2}}{s_{y}^{2}} = r_{xy}^{2} = \frac{\sqrt{r_{12}r_{1\mathrm{II}}r_{\mathrm{I}2}r_{\mathrm{I\,II}}}}{r_{1\mathrm{I}}r_{2\mathrm{II}}}. \tag{15}$$

The index of relevance is

$$\frac{{}_{x}s_{y}}{s_{y}} = r_{xy} = \frac{\sqrt[4]{r_{12}r_{1\mathrm{II}}r_{\mathrm{I}2}r_{\mathrm{I\,II}}}}{\sqrt{r_{1\mathrm{I}}r_{2\mathrm{II}}}}. \tag{16}$$

Both forms of the test must be *parallel,* as must both sets of criterion scores, but *equivalence* is not required in either case.

Upper limits

It is of interest to consider the upper limits of the coefficients and indices of prediction and validity, as either the relevance or the reliability *of the test* approaches unity. We shall consider first the case in which the relevance becomes perfect. The *relevance upper limit* of the *prediction coefficient* will be reached when the test scores and the criterion scores are parallel sets. Then from (9), replacing x_1 by x_2 ($= y_2$), $r_{12}^2 = R_1R_2$: the relevance upper limit of the prediction coefficient is the product of the test and criterion reliability coefficients. Similarly, from (10), $r_{12} = r_1r_2$ under the same conditions: the relevance upper limit of the index of prediction is the product of the test and criterion indices of reliability. For *validity,* we replace y by x in (13) and compare with (11): the relevance upper limit of the validity coefficient is the reliability coefficient of the test. Similarly, from (14) and (12), the relevance upper limit of the index of validity is the index of reliability of the test.

A statement which is widespread in the literature runs to the effect that "the (relevance) upper limit of a test's validity is the square root of its reliability." This statement arises out of a failure to employ exact and consistent definitions to differentiate coefficients from indices. In fact, most previous treatments term r_{1y} the validity coefficient, whereas it is properly the *index* of validity. Also the reliability coefficient is often improperly defined as $r_{1\mathrm{I}}$ rather than as r_{1x}^2. It is then observed that $r_{1\mathrm{I}}$, and r_{1y} are both first-power correlation coefficients, and this leads at once to the erroneous statement given above. As soon as we define all

coefficients as variance ratios and all *indices* as standard deviation ratios, however, it is clear that r_{1y} is an index, whose relevance upper limit is the index of reliability of the test, r_1 (or $\sqrt{R_1}$ or r_{1x}). The equivalent-test correlation, r_{1I}, is a computed estimate of R_1 or r_{1x}^2.

The *reliability upper limit* of the prediction coefficient is found by replacing x_1 by x in (1) and (2). Then, comparing with (13) and (14), we see that the reliability upper limit of a test's prediction coefficient or index is the validity coefficient or index of the criterion, considered as a predictor of the test scores. In the case of validity, we replace x_1 by x in (13) and (14), and compare with (15) and (16): the reliability upper limit of a test's coefficient or index of validity is its coefficient or index of relevance.

Assumptions and experimental limitations

Whenever we introduce an "assumption" or "requirement" in deriving a formula, we imply certain specific limitations in the corresponding structure of the instruments of measurement (tests or criterion scores) or in the conditions of experimental application. Suppose we have two forms of a test, say forms 1 and I, and we proceed to apply them to a group of examinees and compute r_{1I}. If the forms are parallel but unequally reliable, r_{1I} is, by (10), the geometric mean of the reliability coefficients of the two forms. If they are equivalent, r_{1I} is, by (11), the reliability coefficient of each of them. If they are *not* parallel forms, r_{1I} is, by (2), simply the index of prediction of either form by the other.

Suppose we have two sets of items which appear to be essentially equivalent, and arrange them as the odd and even items of a single test. The value of r_{1I} is then the reliability of the test *at the precise time when it was given.* The errors of measurement do not include *quotidian variability:* chance and cyclical fluctuations in the performance levels of the individuals measured or observed and rated. If we use this coefficient in estimating the relevance of the test by (15) or (16), the result will be an underestimation. Quotidian variability will enter into y_2 and y_{II} as error, but it will enter into x_1 and x_I as a b factor which is a part of x, as may be seen by examining (7) and (8). In the case of a criterion, we almost always desire the reliability coefficient to reflect the stability of the *average* performances of the persons observed. We are not interested in the reliability of criterion performances made on June 15 at ten o'clock in the morning, but rather in their average reliability over the period, say, from June 15 to September 15. The two sets of observations on which we are to base our reliability computation should, therefore, be taken sufficiently far apart to permit all types of irregularities in performance to appear or

disappear, and to permit cyclic changes (which may vary in period from one person to another) to reach points such that the levels at the second set of observations are essentially uncorrelated with the levels at the first set of observations. On the other hand, the two sets of observations should not be taken so far apart as to permit any substantial differences in progressive change of performance level to appear. If everybody improves by the same amount in the interval, no harm is done, but if some improve more than others, and particularly if some improve in real performance level while others decline, there will be a spurious reduction in the obtained reliability.

Another source of criterion unreliability appears whenever the observations have to be interpreted or rated. This is observer subjectivity. To avoid correlated subjective errors, leading to spurious overestimation of criterion reliability, two observers should be picked, at random if possible, from a group of presumably competent observers; the first set of observations should be made by one observer and the second set by the other observer. To reduce observer subjectivity, we should use two *groups* of observers, the allocations of observers to the two groups being random. Each observer should make independent interpretations and ratings, and each group of ratings of each person observed should then be averaged to obtain the two sets of criterion scores.

When the observers rate large segments of behavior, or even the total quality of a sample of observations, rather than single acts or short sets of closely related acts, there are further difficulties. In such cases different observers may not notice the same aspects of the acts, they may not even notice the same acts, and they are sure to be unequally impressed by the various features of the behavior. In such cases the criterion scores at which they arrive are essentially nonparallel, and the meaning of the reliability coefficient becomes more or less uncertain. Many of the typical rating scale procedures belong to this class.

A somewhat different type of disturbance arises from the fact that individuals are unequally consistent in their behavior. If we could obtain several independent and strictly equivalent samples of the criterion performance of one person, and could score them all in the same units, we could compute the standard deviation of the sample scores, which would then be the standard error of measurement of one sample score *for him*. The standard error of measurement which is estimated for a group by the usual formulas, based on only two sample scores for each person, is an *average value for the group*, and does not apply strictly to any one person. In practice we cannot readily make studies of the standard error of measurement for each person in a group, both because of the difficulty of obtaining strictly equivalent sets of observations, and because unequal

progressive changes in performance level would occur in most cases before the observations could be finished.

The index of relevance is the test-criterion correlation corrected for attenuation in both, and is given by formula (16). The requirement that *all* the e's be uncorrelated with one another and with both x and y imposes special restrictions on the experimental design. The two forms of the test, x_1 and x_I must be parallel, the two sets of criterion scores, y_2 and y_{II}, must also be parallel, and the experimental conditions must be such that there are no error factors common to x_1 and x_I or y_2 and y_{II} which are not common to all the other pairs of scores. Any such factors will enter into the estimates of x and y, and thereby spuriously reduce the estimate of relevance, r_{xy}. The commonest of such extraneous factors are those arising out of quotidian variability. The nearest we can come to experimental control of these factors is to follow this rule: *The time interval separating the administration of the two forms of the test should equal the average time interval separating the observations of the two samples of criterion behavior.* The time interval separating the average date of administration of the two tests from the average date of the two sets of criterion observations should also be approximately equal to the time intervals separating the two test sessions and the two sets of criterion observations.

Extraneous factors may also be introduced by unequal changes in adaptation to the test situation during the time elapsing between the administration of the first form and the administration of the second. The obvious control is to secure maximum adaptation to the first situation, if necessary by the use of preliminary practice tests.

If in using an essay examination we are trying to measure no more than the effective knowledge of the students at *about* the time examined, we must still consider at least four types of error in estimating the reliability of this test: (1) the particular set of questions as a sample of all possible questions covering the area of effective knowledge to be examined, (2) quotidian variability in the students, (3) permanent biases in the grader, and (4) random errors in grading (snap judgments, for example). Suppose we wish to validate an objective test against such a criterion. How shall we compute r_{2II} so as to include all four of these types of error in e_2 and e_{II} rather than in y? We should first prepare a considerable number of questions, and allocate them to general topics and types. Within each topic and type we should then choose two questions strictly by chance, and allocate one to each form of the essay test criterion. One form of the essay test should be given on one day, and the other

form some days later; the objective test should probably be given at a time about half-way between the times when the two forms of the essay test are given. Better yet, the objective test might be divided into three approximately parallel or equivalent parts; one should be given before the first essay test, one between the two essay tests, and one after the last essay test. Among all available competent graders, we should select *by chance* one (or one group) to grade the first essay test, and another (or another group) to grade the second. Under these conditions we should obtain a reasonably good estimate of the objective test's validity by (13) or (14). This is, of course, subject to the proposition that the answers given by the students to the questions of each form of the essay test may be considered to be a representative demonstration of their effective knowledge of the areas covered by these questions; that is, that the two sets of questions will in fact evoke parallel and perfectly relevant sets of answers.

In the derivation of formulas (13), (14), (15), and (16), it was necessary to assume parallelism, but not equivalence. This point is important: it may be much easier in practice to obtain two parallel sets of criterion scores when there is no requirement that these two sets of scores be equally reliable. There are several alternative formulas (see chapter 15) for r_{1y} and r_{xy}, but all of them require in practice that we obtain *equivalent* pairs of criterion scores and test scores.

Interpretations

Test results are commonly interpreted in terms of standard errors of estimate rather than in terms of variance errors of estimate, and in terms of standard deviation ratios rather than in terms of variance ratios. The standard deviation ratio is in fact the same thing as the standard error of estimate when all performances are expressed in terms of standard scores. The standard error of estimate has the obvious advantage that its units are the score units rather than their squares. It seems reasonable to advocate, therefore, that prediction, validity, relevance, *and reliability* be uniformly reported in terms of indices: r_{12}, r_{1y}, r_{xy}, and r_{1x} ($= \sqrt{R_1}$ or $\sqrt{r_{1I}}$). This will require one general change: reliabilities will need to be reported in terms of the *index of reliability* rather than in terms of the *reliability coefficient.* (Test publishers should welcome this suggestion; the index of reliability always has a higher numerical value than has the reliability coefficient!)

The index of prediction, r_{12}, is ordinarily interpreted in terms of the standard error of estimate, $s_{2.1} = s_2\sqrt{1 - r_{12}^2}$, the coefficient of aliena-

tion, $k_{12} = \sqrt{1 - r_{12}^2}$ (which, as suggested previously, might better be termed the index of nondetermination), and its complement, the index of forecasting efficiency, $E_{12} = 1 - \sqrt{1 - r_{12}^2}$. It is well known, however, that k_{12} remains persistently large, and E_{12} correspondingly small, for all moderate values of r_{12}. Thus if $r_{12} = .866$, $k_{12} = E_{12} = .5$, and even if $r_{12} = .95$, $k_{12} = .31$, and $E_{12} = .69$. In a more usual case, as when $r_{12} = .6$, $k_{12} = .8$, and $E_{12} = .2$; the standard error of estimate is only 20 percent smaller than the standard deviation of the criterion scores. This method of interpretation is the one to be used when the criterion performances of particular individuals are to be predicted from their test scores. If we are using the test for *selection,* however, we may prefer a different interpretation. The best possible selection would be selection on the basis of the criterion performance itself. Suppose our present group of N has a mean criterion score M. If we could select the best n out of the N, say, the mean criterion score of the n selected would be M'. If instead we select the best n out of N on the basis of their *test* scores, the mean criterion score of *these n* will be M_T. Brogden (*1*) has shown that in this case,

$$r_{12} = \frac{M_T - M}{M' - M}. \tag{17}$$

The numerical value of the test-criterion correlation is the ratio of the mean criterion improvement effected by test selection, to the mean criterion improvement that could have been secured by selection on the basis of the criterion scores themselves. The computation of r_{12} is based on the total group of N.

Validity and test reliability

Criterion performances are usually the best we can afford to obtain as regards their reliability, after taking all possible care to insure maximum relevance. We can, in fact, validate a test against any perfectly relevant criterion, provided its reliability is appreciably higher than zero and can be determined with sufficient accuracy. The more reliable the criterion, however, the smaller in general may be the sample used in estimating its reliability and in estimating the validity or relevance of the test. The standard error of any correlation coefficient, for example, $r_{2\text{II}}$, computed on the basis of a sample of N cases, is $(1 - \rho_{2\text{II}}^2)/\sqrt{N-1}$, where $\rho_{2\text{II}}$ is the corresponding correlation in the general population, from which the sample of N may be considered a *random* sample. Its value can be reduced either by increasing $\rho_{2\text{II}}$ or by increasing N.

Tests, and particularly indirect tests, are usually made as short as possible, consistent with adequate validity and predictive power. The *relevance* of a test, however, is not a function of its reliability, aside from considerations analogous to those of the preceding paragraph. If we have two equivalent forms of a test whose standard deviations are equal, and have given them to a group for whom we have a set of relevant criterion scores (or two parallel sets of such scores), we can estimate the predictive power or validity of a longer or shorter test, or the amount of lengthening or shortening necessary to obtain some specified level of predictive power or validity.

Let the test scores be x_1 and x_{I}, the criterion scores y_2 and y_{II}, and n the amount of lengthening or shortening, expressed as a multiple or fraction of x_1 or x_{I} (which must be of equal length if they are parallel, equally reliable, and equally variable). Then,

$$r_{n2} = r_{x2}\sqrt{R_n} = r_{x2}r_n. \tag{18}$$

$$r_n = \sqrt{R_n} = \frac{r_{n2}}{r_{x2}}. \tag{19}$$

$$r_{ny} = r_{xy}\sqrt{R_n} = r_{xy}r_n. \tag{20}$$

$$r_n = \sqrt{R_n} = \frac{r_{xy}}{r_{ny}}. \tag{21}$$

Formulas (18) and (19) relate to predictive power and require only one set of criterion scores; (20) and (21) relate to validity and require two parallel sets of criterion scores. All four of the formulas require two equivalent and equally variable (and hence equally long) sets of test scores in order to estimate R_n, the reliability coefficient of the lengthened test, and its square root, r_n, the index of reliability of the lengthened test. Formula (18) gives the index of prediction of the lengthened test, and formula (20), the index of validity of the lengthened test. In these formulas r_{x2} is analogous to the r_{1y} of formula (14):

$$r_{x2} = \sqrt{\frac{r_{12}r_{\mathrm{I}2}}{r_{1\mathrm{I}}}}\,. \tag{22}$$

Also r_{xy} is the index of relevance as given by (16), and R_n is obtained by the Spearman-Brown formula,

$$R_n = \frac{nr_{1\mathrm{I}}}{1 + (n-1)r_{1\mathrm{I}}}\,. \tag{23}$$

To estimate the amount of lengthening or shortening necessary to obtain some specified index of prediction, r_{n2}, or some specified index of validity, r_{ny}, we substitute the specified value in (19) or (21) and solve for r_n, square the value of r_n to obtain R_n, and find the value of n by the inverted Spearman-Brown formula:

$$n = \frac{R_n(1 - r_{11})}{r_{11}(1 - R_n)} \cdot \tag{24}$$

In all of the lengthening procedures it is assumed that the added items will be similar in the distributions of their difficulties and discriminations to those of the tests and criterion scores on which the computations are based. In shortening a test, we typically retain the best items rather than a strictly random sample. By this procedure we are often able to shorten a test with considerably less loss of reliability, predictive power, and validity than the formulas would imply.

Predictive power under conditions of restriction and curtailment

The numerical values of correlation coefficients depend upon the variabilities of the groups measured, as well as upon the intrinsic strengths of the relationships involved. If two tests have been correlated with essentially the same criterion, but on different groups, we should compare the respective *standard errors of estimate of the criterion scores,* rather than the test-criterion correlation coefficients, to find out which test is the better indicator of the criterion performances.

It is often necessary to carry out a prediction study using a restricted or curtailed group. When this has been done, we may wish to estimate what the relationship would have been in the unrestricted or uncurtailed group. Often we know the standard deviation of at least one of the variables in the unrestricted or uncurtailed group. We can make such estimates, subject to one basic assumption: namely, that the essential effectiveness of the predictor is the same in the two groups, which implies that the standard error of estimate is the same. Let the test or predictor be variable 1, the criterion variable 2, and the variable on the basis of which the restriction or curtailment took place (if other than variables 1 or 2), variable 3. Let R and S denote correlations and standard deviations in the unrestricted or uncurtailed group (the former estimated; the latter, if present in the formulas, computed), and let r and s denote the correlations and standard deviations computed for the restricted or curtailed group.

In one common type of situation we may wish to estimate the predictive power of a test in a group for which no criterion measures are available, by reference to data based on another similar group. The standard errors of estimate are assumed to be the same in the two groups, but the stand-

ard deviations are not. The data are the test standard deviation and the test-criterion correlation in the group on which the prediction study was based (s_1 and r_{12}), and the test standard deviation in the new group (S_1). The estimated test-criterion correlation or index of prediction in the new group is

$$R_{12} = \sqrt{1 - s_1^2\left(\frac{1 - r_{12}^2}{S_1^2}\right)}\ . \tag{25}$$

Another common situation is one in which the predictor is used for selection purposes on the basis of its assumed predictive power before this has been determined. We can compute the test-criterion correlation for the selected group only, but wish to estimate its value for the entire applicant group. We know the standard deviation of the test (predictor) for the applicant group, but not the standard deviation of the criterion. The estimate in this case is:

$$R_{12} = \frac{\dfrac{r_{12}S_1}{s_1}}{\sqrt{1 - r_{12}^2 + \dfrac{r_{12}^2 S_1^2}{s_1^2}}}\,. \tag{26}$$

In still another typical situation we administer variable 1 to a group which has already been selected on the basis of variable 3. We know S_3, but we do not know either S_1 or S_2. The selector, variable 3, might be high school grade-point average; the new predictor, variable 1, some scholastic aptitude test, and the criterion, variable 2, the college freshman grade-point average. The test is assumed to have been administered only to those who were admitted to college. We wish to estimate the test-criterion correlation for the total applicant group.

$$R_{12} = \frac{r_{12} + r_{13}r_{23}\left(\dfrac{S_3^2}{s_3^2} - 1\right)}{\sqrt{\left[1 + r_{13}^2\left(\dfrac{S_3^2}{s_3^2} - 1\right)\right]\left[1 + r_{23}^2\left(\dfrac{S_3^2}{s_3^2} - 1\right)\right]}}\,. \tag{27}$$

If the new predictor is given to the total applicant group, we will know not only S_3, but also R_{13}. In this case a better estimate is given by

$$R_{12} = \frac{r_{12}\sqrt{1 + R_{13}^2\left(\dfrac{s_3^2}{S_3^2} - 1\right)} + R_{13}r_{23}\left(\dfrac{S_3}{s_3} - \dfrac{s_3}{S_3}\right)}{\sqrt{1 + r_{23}^2\left(\dfrac{S_3^2}{s_3^2} - 1\right)}}\,. \tag{28}$$

These formulas require the use of product-moment correlations throughout. If variable 1, the predictor, is continuous, and variable 2, the criterion, is arbitrarily dichotomized (pass-fail, graduated-resigned, and the like), we cannot substitute biserial correlations in these formulas, because if the original distribution is normal in variable 2, the restricted distribution is not likely to be normal also.

Multiple correlation techniques

It is doubtful that any other statistical techniques have been so generally and widely misused and misinterpreted in educational research as have those of multiple correlation. This is not wholly nor even mainly the fault of the educational research workers. The texts—including even the most recent and advanced mathematical treatments—are quite generally silent or unclear concerning the experimental limitations and interpretational restrictions implied by the assumptions on which the formulas have been derived. Correct formulas for the solution of most of the problems for which multiple correlation and regression techniques are incorrectly used in educational research have never been derived. In order to emphasize these points we shall state them first without discussion and in the strongest terms, with particular reference to the weighting of subtests in a battery and the estimation of the predictive power of the resultant weighted battery scores.

a) Only in exceptional cases are the multiple regression coefficients of a criterion score upon a set of test scores the proper weights to give the test scores in order to predict or estimate the criterion scores.

b) When the statistics from a given sample have been used to determine the test score weights, the estimate of the aggregate or multiple correlation of the tests with the criterion, as computed from the data of that sample, *is not an estimate* of the predictive power or validity of the battery.

c) If items from an experimental test are selected and/or weighted on the basis of item analysis data from a given sample, that sample cannot be used to determine the weight of the test composed of the selected and/or weighted items, relative to any other test or tests; nor can it be used to determine the predictive power or the validity of the test composed of the selected and/or weighted items.

Regression coefficients and linear restraints

It is difficult, if not impossible, to assemble a battery of indirect tests to predict or estimate a set of criterion scores, without finding that some

or most of the predictors correlate fairly highly with some or most of the others. In the language of factor analysis, the number of tests exceeds the number of factors measured by these tests. We can reproduce all the test scores, within the limits of sampling error, by the use of a set of factor scores smaller in number than the test scores. We cannot disprove the hypothesis that some of the test scores are merely linear combinations of some of the others, or that all of the test scores are linear combinations of some smaller number of underlying and essentially independent traits or abilities. The test system is said to possess as many *linear restraints* as the difference between the number of tests and the number of traits or abilities which the tests can measure with greater than chance certainty. If we attempt to determine multiple regression coefficients for all the tests, some of these coefficients and perhaps all of them will assume values that are essentially the effects of chance.

A battery of four tests was given to two groups of mechanics. The jobs were appreciably different, but both involved highly skilled technical work. The criteria were supervisors' ratings, the same rating scale (a single over-all rating) being used with both groups. The criterion correlations were as follows:

Test	1	2	3	4	(N)
Group I......	.65	.60	.71	.47	(51)
Group II......	.36	.76	.72	.54	(84)

In group I, r_{23} was .66, this being the second highest intercorrelation among the tests. In group II, r_{23} was .76, the highest intercorrelation. In group I, test 3 had the highest beta weight; the beta weight for test 2 was *low negative*. In group II, test 2 had the highest beta weight; the beta weight for test 3 was positive but just above zero. Actually tests 2 and 3 were about equally valid in both cases. Whichever of the two happened to have the higher criterion correlation became the most heavily weighted test in the battery, while the other acquired a weight close to zero. To obtain the best prediction, we should not discard test 2 for group I and test 3 for group II. We should weight them equally and treat them as a single variable in both regression equations.

In more complex cases, where one predictor is essentially a linear combination of several others, or where several or all of the predictors are linear combinations of a smaller number of underlying traits or abilities, the solution becomes exceedingly complex. A fairly complete analysis of this problem has been given by Davis (3), but the technical difficulties of the treatment will be too great for those who do not have extensive training in advanced mathematical statistics.

Multiple regression and multiple correlation

After we have reduced our battery, either by eliminating tests or combining tests, to the point where linear restraints are no longer present, we may determine the "best" weights by the multiple regression procedure. In doing so, we are finding the weights which are "best" for the given sample. The least squares procedure gives the best possible fit to the sample statistics, including their sampling errors. The multiple correlation in the sample is as high as it can be for that given set of individuals. If we draw a second sample of persons from the same population, give them the same tests, and secure similar criterion scores, we shall have to use the regression coefficients from this second sample to obtain a multiple correlation coefficient which is equally likely to be above or below the one obtained from the first sample. If we apply the regression weights from the first sample to the scores from the second, the resulting aggregate correlation in the second sample will necessarily be lower. But this is exactly what we do when we use a test battery for any practical purpose. We determine the regression weights on the basis of statistics from an experimental sample. Then we use them in making predictions concerning the criterion behavior of *other* persons from the same population.

The predictive power or validity of a battery must, therefore, be determined by giving it to a second sample from the same population as the sample used in determining the weights. Using the weights from the first sample, we determine the aggregate correlation in the second sample. This will almost always be substantially lower than the multiple correlation in the first sample, and it will necessarily be lower than the multiple correlation in the second sample. This aggregate correlation, and not the multiple correlation in either sample, is the index of prediction for the battery. When divided by the index of reliability of the criterion from the second sample, it becomes the index of validity of the battery. This procedure is termed "cross-validation." It is the only procedure which provides an uninflated estimate of the predictive power or validity of a weighted battery.

Item analysis, test selection, and validity

These same considerations apply with even greater force when we are dealing with single items instead of tests. Whether the procedure be item selection or item weighting or both, we are dealing partly with real relationships and partly with chance. The selected items are the ones which are best in the given sample. The item weights are also the ones which are best in that sample. The validity of the reduced and/or weighted test

must therefore be determined on the basis of data from another sample. A "valid test" has in fact been constructed by purely chance methods. Every item score of every person "tested" was obtained by a method equivalent to flipping a coin and recording "right" if it fell heads. By selecting the "best" 24 items out of 85, using a sample of 29 students, with college grade-point average as the criterion, a correlation of .82 was obtained between the "test scores" and the grade-point averages of the same 29 students! The reliability of the "test," quite properly, was zero. The smaller the sample, and the smaller the proportion of items selected, the greater will be this spurious validity or predictive power.

These same principles apply in the case of test selection. If we give a large experimental battery to a sample of individuals, and select, say, the best three or four tests out of ten or fifteen, we cannot estimate the battery validity or predictive power on the basis of data from that sample, whether or not we also use the same data to weight the tests.

We need one sample for item selection and item weighting. We need another for test selection and test weighting; and the latter process, to be effective, is much more complicated than the grinding out of a set of multiple regression coefficients. We need a third sample to determine the predictive power or the validity of the battery. Every time we violate any one of these rules, we increase spuriously the apparent validity or predictive power of our test or battery. If we plead practical necessity, due to limited availability of experimental subjects, we are in even worse straits. With small samples the spurious elements increase enormously. Words of caution concerning interpretation are not a substitute in such cases for sound research.

Selected References

References Cited in the Text

1. Brogden, H. E. "On the Interpretation of the Correlation Coefficient as a Measure of Predictive Efficiency," *Journal of Educational Psychology*, 37: 65–76, 1946.
2. *Critical Requirements for Research Personnel.* ("American Institute for Research, Research Notes," No. 2.) Pittsburgh: The Institute, June 1949. 4 pp.
3. Davis, H. T. *The Analysis of Economic Time Series.* ("Cowles Commission for Research in Economics," Monograph No. 6.) Bloomington, Ind.: Principia Press, Inc., 1941. Pp. 97–101 and chap. 5.
4. *The Development of a Procedure for Evaluating Officers in the United States Air Force.* ("American Institute for Research, Research Notes," No. 1.) Pittsburgh: The Institute, June 1949. 4 pp.
5. Flanagan, J. C. "A New Approach to Evaluating Personnel," *Personnel*, 26: 35–42, 1949.
6. ———. "Job Requirements." In *Current Trends in Industrial Psychology.* Pittsburgh: University of Pittsburgh Press, 1949. Pp. 32–54.
7. Mosier, C. I. "A Critical Examination of the Concepts of Face Validity," *Educational and Psychological Measurement*, 7: 191–205, 1947.

Further References on the Meaning of Validity and on the Statistics of Estimation

8. Cronbach, L. J. "Response Sets and Test Validity," *Educational and Psychological Measurement,* **6**: 475–94, 1946.
9. Gillman, L., and Goode, H. "An Estimate of the Correlation Coefficient of a Bivariate Normal Population When *X* Is Truncated and *Y* Is Dichotomized," *Harvard Educational Review,* **16**: 52–55, 1946.
10. Guilford, J. P. "New Standards for Test Evaluation," *Educational and Psychological Measurement,* **6**: 427–38, 1946.
11. Jarrett, R. F. "Percent Increase in Output of Selected Personnel as an Index of Test Efficiency," *Journal of Applied Psychology,* **32**: 135–45, 1948.
12. Kelley, T. L. *Interpretation of Educational Measurements.* Yonkers-on-Hudson, N.Y.: World Book Co., 1927. Pp. 171–85.
13. Richardson, M. W. "The Interpretation of a Test Validity Coefficient in Terms of Increased Efficiency of a Selected Group of Personnel," *Psychometrika,* **9**: 245–48, 1944.
14. Rulon, P. J. "Validity of Educational Tests." In *National Projects in Educational Measurement,* ed., K. W. Vaughn. Washington: American Council on Education, 1947. Pp. 13–20.

References on Industrial and Military Criteria

15. Bechtoldt, H. P. "Problems in Establishing Criterion Measures." In *Personnel Research and Test Development in the Bureau of Naval Personnel,* ed., D. B. Stuit. Princeton, N.J.: Princeton University Press, 1947. Chap. 19.
16. Bellows, R. M. "Procedures for Evaluating Vocational Criteria," *Journal of Applied Psychology,* **15**: 449–513, 1941.
17. Hoppock, R. (ed.). "Criteria of Vocational Success—A Symposium," *Occupations,* **14**: 917–75, 1936.
18. Horst, P. *The Prediction of Personal Adjustment.* ("Social Science Research Council Bulletin," No. 48.) New York: The Council, 1941. Chap. 3.
19. Jenkins, J. G. "Validity for What?" *Journal of Consulting Psychology,* **10**: 93–98, 1946.
20. Patterson, C. H. "On the Problem of the Criterion in Prediction Studies," *Journal of Consulting Psychology,* **10**: 277–80, 1946.
21. Stuit, D. B. "The Effect of the Nature of the Criterion upon the Validity of Aptitude Tests," *Educational and Psychological Measurement,* **7**: 671–76, 1947.
22. Taylor, E. K., and Tajen, C. "Selection for Training: Tabulating Equipment Operators," *Personnel Psychology,* **1**: 341–48, 1948.
23. Thorndike, R. L. *Personnel Selection: Test and Measurement Techniques.* New York: John Wiley & Sons, Inc., 1949. Esp. chap. 5.
24. Toops, H. A. "The Criterion," *Educational and Psychological Measurement,* **4**: 271-97, 1944.

17. Units, Scores, and Norms

By JOHN C. FLANAGAN
University of Pittsburgh

COLLABORATORS: Edward E. Cureton, *University of Tennessee;* Walter N. Durost, *Boston University;* Eric Gardner, *Syracuse University;* H. A. Greene, *State University of Iowa;* Harold O. Gulliksen, *Princeton University;* Paul Horst, *University of Washington;* Truman L. Kelley, *Harvard University;* L. L. Thurstone, *University of Chicago*

The Problem of Interpreting Test Scores

TEST SCORES ARE MEANINGFUL AND VALUABLE TO THE EXTENT THAT they can be interpreted in terms of capacities, abilities, and accomplishments of educational significance. Interpretation thus presents the problem of establishing procedures suitable for describing, recording, and comparing the performances of individuals in specific test situations. Since the point of view of this discussion involves the adoption of a rather broad definition of a test, this problem becomes one of describing any observation of an individual's behavior in a defined situation. The more detailed and specific the description is, the better, within practical limits. While very comprehensive descriptions can be defended on theoretical grounds, they sometimes become too cumbersome and complicated for practical use. A practical reason for adopting a simplified descriptive procedure in reporting test behavior is to provide a set of symbols which may be readily manipulated and understood. A long verbal record of responses to a series of test situations may be completely descriptive, but it is often quite difficult to communicate, compare, or combine with other observations.

This discussion is concerned primarily with the problem of assigning numerical values to correspond to behavior samples which are sufficiently restricted, defined, and controlled to permit general agreement by judges or scorers as to the characteristics of the behavior. Such observations are called "objective" to distinguish them from those substantially dependent on the personal judgment of the observer and designated as "subjective."

There are many aspects of test behavior that may be selected for observation. Naturally it is not practical to attempt to observe and describe all phases of test performance, but, in general, the greater the number of

characteristics observed, the more accurate the description of the test performance reported can be. There are limits beyond which it is impractical to go in the reduction or the extension of the number of categories to be reported. Another important question involves the determination of what description will be most useful. In some cases a classification of results as "satisfactory-unsatisfactory" may be adequate. In others a finer discrimination among the members of the group may be required. The actual refinement of the discrimination ultimately depends upon the nature of the use to be made of the results.

A test score is a number assigned to the performance of an individual to define the relative value of his accomplishment. As a means of expressing and recording observations of behavior, numbers have many advantages, among which are simplicity of communication, comparison, and manipulation for interpretational purposes. It should be recognized that it is possible to oversimplify the numerical expression of achievement. In describing test results, several numbers or scores almost always will provide more information than a single number. It is more descriptive, for the usual time limit tests, to have both the number of questions answered correctly and the number answered incorrectly than a number made up of a combination of these, granting that a valid combination of these factors could be made. A much more complete description would include the designation of the particular questions which were answered correctly and those which were answered incorrectly. In most classroom situations further information regarding the nature of the incorrect answers would be relevant both for the improvement of the testing instrument and for the improvement of instruction following its use.

Educational tests almost universally may be regarded as samplings from a much larger field of knowledge, information, and skill. While the extent and nature of the sampling may have an important effect on the utility of the test itself, the user of the test results is primarily interested in the inferences he can make from the results rather than in the description of the test behavior itself. For example, in a word-meaning test it is not so important to know that the individual selected the correct synonym for a specific test word. The important factor here is the accuracy of the inference that can be made regarding his knowledge of words of the types sampled by the test. In some instances it is desired to make broader inferences from such results concerning the individual's ability to learn the meanings of such words, to learn other things, or to understand what he reads. Furthermore, the single number which best indicates the individual's knowledge of words might not be the same as the numbers which provide

the best basis for other types of inferences. In all practical test situations several numbers (scores) can be shown to be more informative than a single score. The choice of scores to be reported must be decided in terms of the practical use to be made of the test results in each specific case.

For most homogeneous tests, reliance is placed on a single number which is called the individual's score on the test. Such a score could be obtained in a number of ways. In the evaluation of products which cannot be treated as right or wrong, as is the case with samples of compositions, handwriting, freehand drawing or lettering, sewing, etc., product scales expressing quality in terms of arbitrarily defined judgment units are used. Several different techniques for constructing such scales have been proposed, but in general they are based upon the determination of the proportion of judges who recognize degrees of quality of samples representing a wide range of performance with respect to the characteristics in question. In spite of the fact that this procedure utilizes judgments of questionable reliability, it does result in the arrangement of quality specimens in an order-of-merit scale which has practical possibilities.

In the teacher-made examination of the essay type the score is a subjective estimate of the degree to which the pupil's response attains a quality level similar to one of the reference levels in the mind of the rater. The difficulty with this type of score lies primarily in its unreliability. Standards vary from person to person and from time to time; that is, the quality levels in the minds of the raters are not the same from one day to the next nor from one rater to the next. In contrast, the score for the objective-type examinations is primarily based on the number of items answered correctly, and the scores are to a large degree independent of when or by whom the rating is done.

In fixed-choice situations such as alternate- or multiple-response items the test score is usually adjusted if some individuals do not mark all items in the test. This correction for "guessing" is based on the number of choices per item and the number of wrong choices selected.

A related problem of scoring is the problem of the amount of credit to be given for a particular question. The advantage of giving varying amounts of credit for items differing in difficulty or importance is usually not of great practical importance in educational measurement. If a topic is deserving of greater weight in the total sampling, this usually can be accomplished by allotting more items to that topic.

It has been observed that for practical use simple numbers are desirable. For purposes of interpretation it is also desirable that these numbers be as directly meaningful as possible. As noted above, the test user is generally

much more interested in the inferences to be made from the test scores than in the actual number of items correctly answered. There is, therefore, no particular reason for preserving the test score in terms of the number of correct responses. This has led to the selection of certain numbers as basic points of reference and also to the empirical establishment of meaningful units of measurement.

In defining basic points of reference for score scales, several points have usually been considered. One of these has been the average score of a specified group having received a defined amount of training. These descriptive numbers are usually referred to as "norms." Typical norms would be the average score obtained by members of classes having received two years of instruction on the French language or the average score of those having just completed third-semester algebra.

There are many difficult problems involved in the establishing of test norms of all types. The results from schools of various types and in different communities differ much more than is generally known. Different textbooks, study plans, and methods, as well as different instructional personnel, all affect results. The aptitude of students for the specific type of material, the aims of instruction, and the quality and amount of instruction influence test results. It has become apparent that there is no such individual as a "fifth-grade pupil," nor a "twelve-year-old." The twelve-year-old child in the seventh grade is quite a different person, having had many additional types of experiences, from the twelve-year-old who is in the fourth grade. Thus, the meaning and the significance of grade norms and even age norms. without more exact definition than is usually given, are both coming under serious questioning.

Test "norms" may be defined as estimates of some characteristic of a distribution of test scores for a specified population. Norms describe the actual performance of specified groups of individuals. "Standards," on the other hand, are *desirable,* or desired, levels of attainment, preferably expressed in terms of outcomes of instruction. For example, the norm of quality of handwriting at the end of the fifth grade is 50 on the Ayres Handwriting Scale. On the other hand, studies indicate that a satisfactory standard of legibility on this same scale is approximately 70. The standard for arithmetic accuracy in routine computation is 100 percent. The norm for such accuracy for the eighth-grade students in a particular school system may be 88.6 percent. The score of an individual as obtained on a French reading test might be at the tenth-grade norm. This gives little information about how well he reads various types of materials. The probable degree of comprehension of the individual in reading a typical French

newspaper would provide a useful social standard for interpreting scores on a French reading test.

A further desirable characteristic of test scores may be designated as comparability. Scores (either raw or derived) on two or more tests may be said to be *comparable* for a certain population if they show identical distributions of "true" scores for that population. Scores on the same tests, of course, may be comparable for one population and not comparable for another. Comparability can refer either to the scores from several forms of the same test or to scores from tests of different abilities. Different forms of the same test must yield comparable scores if they are to be completely interchangeable in use. Such interchangeable forms of a single test that yield comparable scores are called *equivalent* forms.

In highly refined instruments the equivalent forms of the tests are not only closely equated in the process of construction, but they are also accompanied by standard score scales which bring about more precise comparability. If results from several tests of different types are to be compared, as in preparing a profile of the results from a battery of tests taken by an individual, it is necessary that the scores from the various tests have some type of meaningful comparability.

The Principal Types of Descriptive Information Derived from Tests

Educational and psychological measurement is essentially a means of obtaining descriptive information about the performance of individuals. The basic goal of testing is to obtain and use this information in an efficient and reasonably precise manner. The broad purposes for which educational test scores are used—educational evaluation and educational guidance—are discussed in the final section of this chapter. Only a brief general statement will be given here. For purposes of evaluation, information is needed that will assist in placing a value on past experiences. Evaluation thus provides an effective basis for the improvement of teaching and learning. For guidance, information is needed that will make possible predictions of the individual's effectiveness in various types of activities in which he may engage in the future.

To fulfill these two major purposes, five main types of descriptive information need to be provided. These include (1) descriptive information in terms of content, (2) rank in a specified group, (3) level of educational development, (4) growth or progress during a fixed period, and (5) the profile or pattern of an individual's abilities in various fields.

Content

The most basic type of descriptive information obtained from tests refers to the individual's knowledge and ability with respect to the content itself. This information tells us directly what the individual did with respect to the questions and problems set by the test. It contrasts with the other types of information in which the individual's score is described by comparison with other scores obtained on the same test. Usually these are scores made by others, but they may also be scores made by the same individual at previous times. Examples of descriptive information in terms of content include such statements as, "This individual knows all of the single digit pairs of addition combinations," or "This individual can spell correctly 90 percent of the words on a specified eighth-grade spelling list," or "This individual knows the English meanings of 4,000 of the 5,000 most frequently used French words," or "This individual can multiply randomly selected three-digit numbers by randomly selected two-digit numbers at the rate of three per minute for a fifteen-minute period with fewer than three errors in the period," or "This individual can translate a typical section of 100 lines from a French newspaper with fewer than five errors."

All of the above statements are inferences based on a specific sample of what the individual did at a certain time. The individual's most probable status with respect to all of the content from which the sample was drawn or his most probable performances on repeated tests can be inferred. With reference to the subsequent discussion of units it should be observed that many of the examples given above are concerned with counting or enumerating, in which case they involve cardinal rather than ordinal numbers. The scoring or counting process ignores differences in the importance, difficulty, or representativeness of items in the test. The statement with respect to the ability to translate materials from a French newspaper provides a practical report regarding performance at a specified level of difficulty. Such statements have direct meaning without reference to the scores of other individuals. This type of information lends itself especially well to setting standards of achievement. Some writers have used the terms "norms" and "standards" interchangeably. This usage appears undesirable in that the word "standard" implies something set up as a desirable model or minimum goal, whereas "norm" carries the connotation of describing things as they actually are. It seems unwise to encourage a usage which tends without criticism and automatically to establish the present average performance of individuals on a test (the "norm") as the acceptable score (the "standard") for that test. Examples of typical norms and standards were given on pages 698–99.

Ranks

The traditional use of tests has been to place people in order with respect to other members of the particular group being tested. The knowledge that an individual's score on a particular test was 63 in itself is not of any significance. If the additional information is obtained that the score is computed by counting the number of items correctly answered and also that there were 90 items of a specified type in the test, it is possible to begin to attach some meaning to the result. For example, it can be said that the individual's score in terms of percent of items correctly answered was 70. In terms of content a definite score has been obtained which may be related directly to standards established by the examiner who prepared the questions. It may be regarded by this examiner as a good or poor score on the basis of his judgments of the difficulties of the items and the expected performance of those taking the examination.

In many cases such judgments are difficult to make and are not related to the realities of the situation. Frequently, when teachers, civil service examining boards, or similar groups discover that all of those examined have fallen below the standard they originally set for the examination, they revise their judgments regarding the content of the test materials and either apply a correction factor to the scores or modify the standards in terms of the obtained scores. The results of applying the test to the group become an important factor in describing the performance of an individual member of the group in many situations. Probably the most important bit of information regarding the group is its mean or average score. Knowing that the mean number of items correctly answered by the group is 56 certainly helps in evaluating the score of 63. Adding information as to the range of scores (the highest was 76 and the lowest, 32) is also useful. A more stable measure of variability is the standard deviation. If the distribution is approximately normal, this provides a fairly adequate basis for placing any score with respect to the score distribution of the group.

Another method of placing the individual's score with respect to those in the group is in terms of rank in the group, or more commonly in terms of centile ranks or in relation to percentile norms. Centile ranks and percentile norms are discussed in detail on pages 717–21 following.

Level of development

Information concerning rank in the group is very valuable in supplementing direct information regarding content. For many purposes these types of information are inadequate. Knowing that an individual has mastered a specified set of skills or items of information is useful. It is

also important to know that his score ranks at the 92nd centile in his class.

Other questions frequently arise if a broad view of the significance of this score is being taken. One of the most important of these questions is that of where the individual stands with respect to other groups. What level of development has he reached? The terms in which level of development is usually expressed are the age, grade, or number of years of study required by the average student to achieve the level of development corresponding to the specified score.

In addition to knowing that a student is at the 94th centile rank in his class, it is certainly useful to have information concerning level of development in terms of school groups, such as that an individual's score on a test of English expression places him at the college senior level. It is also desirable to have information in terms of *content,* such as that he can spell correctly 95 percent of the most common words used in writing or that he can discriminate between the shades of meaning of 5,000 out of 10,000 synonyms whose relations are discussed in a specified dictionary. This takes on even more meaning if we are told that the test score indicates that the individual can write a report of a news event that the typical big city newspaper editor will accept without more than one or two minor corrections in expression. Such information on *level of development* in terms of *content* has important practical implications.

The employer wishes to know that the typing test score made by an applicant for a secretarial position places her at the 80th centile rank for individuals tested at the time of completing a two-year secretarial training course in one of the better schools. He also finds it valuable to know that her corrected score in terms of words per minute for a ten-minute typing test of typical general correspondence material is 65 words per minute. However, the most useful information to him might be that her typing score placed her at the average level achieved by the satisfactory secretaries of executives in his company and similar companies. The student in a typing course also wants to know where she stands, not only with respect to words per minute and the rest of the class, but also with respect to various levels of employed typists.

In studying a foreign language, information regarding level of development in terms of the amount of study required by the average individual to reach that degree of proficiency is regarded as useful. Students are frequently placed in college classes in accordance with such information.

For some purposes, information regarding level of development might be in terms of the average score achieved by a group of individuals in a particular type of work who are beginners or who have one, two, three, five, or ten years of experience, rather than in terms of average scores for

those with particular position titles. It is clear that knowledge regarding level of development expressed in relation to amount of study experience, amount of job experience, or practical job performance provides very useful information for the interpretation of scores on educational tests.

Growth

Any dynamic view of the educational process must take into account not only the status of the individual now, but also the rate at which he is progressing and other characteristics of the curves indicating his growth with respect to various educational areas during a period of time. Knowledge of progress is an important source of motivation for the individual. It is also essential for the planning and administration of an educational program. Growth information should be an important part of any plan for interpreting test scores.

Information regarding growth differs from information regarding level of development in that any measure of growth must involve the comparison of at least two test scores given at the beginning and end of a specified time interval. In order to provide the most valid basis for the comparison of such scores for this purpose, they must be expressed as "comparable" scores obtained from parallel forms of a test of the ability in question. Scores on two tests are comparable for a given population if the corresponding true scores show identical distributions for that population. The technical problems and the serious practical difficulties met in establishing comparable scores are discussed later in this chapter (pages 750–60). In spite of these practical difficulties, comparability can be approximated within reasonable limits if appropriate methods are used in constructing and equating the examinations.

An important factor which affects scores being used to obtain measures of growth is "practice effect." This is the term used to describe the increase in the individual's score which is the direct result of previous experience with the same or a similar form of the test. Practice effect varies with the time interval between the two testings, the nature of the examinations, and the previous test experience of the individuals involved. Tests in which speed is important or tests in which learning during the test period itself is relatively large are especially affected by previous experience. The increment due to experience with a speeded test is especially large when the time required for reading directions and samples is included within the working time on the test. The individual who is able to skip such materials and begin immediately on the test items themselves has an important time advantage. If time is not important, if learning during the test period is negligible, if the materials in the various forms of the

test used are not too nearly identical, and if the individuals have had some previous experience with tests of the same general type, the practice effect from one form to another will not be a serious disturbing factor in studying growth.

In all types of interpretation of test scores there is the problem of how accurately an individual's score reflects the average score which he would obtain if he took an indefinitely large number of parallel forms of the same test. For studying growth, this problem is doubly important since any measure of growth must be based on at least two test performances, each of which involves sampling errors with respect to content and typicality of performance of the individual himself. The difference score therefore has a larger sampling error than a single score, since the two errors can be in opposite directions, and it is important that the magnitude of this error be known and considered in interpreting the results.

Profiles

In planning an instructional or remedial program for the individual pupil or in counseling him concerning educational and vocational plans, it is very important to have accurate information concerning the individual's test performances in a number of different fields. It is clear that in order to compare scores from one subject-matter field to another it is not sufficient to have comparability merely from one form of a test to another; there must be comparability of a score on a test in one field to a score on a test in another field. This type of comparability is very much more difficult to secure than equality for the various forms of a single test. The comparability must refer to the situation as it exists in a clearly defined group, and there are much greater possibilities for lack of uniformity in results even when the defined group is quite large.

A useful method of comparing performance on a number of tests is by means of a graphic record form called a profile, in which some type of equating of the test scales of the different tests has been effected. Grade score equivalents and centile ranks quite frequently are plotted on such profiles, even in many cases where these scores and ranks are based on very dissimilar norm groups and are therefore known to be not comparable. This practice tends to encourage misinterpretation of the graphic records, since the plotting of the results on a common scale suggests a comparability which does not actually exist. For example, an 80th centile rank in botany and a 40th centile rank in Latin for a given student plotted on a graph would certainly encourage the interpretation of marked superiority in science subjects as contrasted with languages. However, if somewhat-below-average students select botany and only unusually able students

choose Latin, it may be that if all students had taken both subjects this same individual would have received a higher centile rank in Latin than in botany. This difficulty could be taken care of by proper adjustments on the profile if the necessary data were available. Differences in the quality and content of instruction from school to school are also important.

The use of derived scores having a common scale (such as the various systems of derived scores discussed later in this chapter) tends to make the scores comparable with reference to a broad, well-defined, basic group. It must be kept in mind that in interpreting the results of such profiles any variation of the individual's experiences from those of the basic group will affect the scores and may complicate, if not invalidate, the comparison of his results on tests in various fields. In spite of these limitations, the importance of self-analysis based on this type of information for guidance and planning purposes is so great that every effort should be made to secure profile data in as meaningful and unambiguous form as possible.

Types of Scores and Norms for Educational Tests

Raw scores

The raw score of an individual on a test may be defined as the number of items to which he responds correctly. This may be modified, in the case of multiple-choice tests, by applying a correction for "guessing" involving the number of available choices and the number of incorrect responses. The problem of correction for guessing is discussed in chapter 10 (pages 347–51). The raw score is a very fundamental piece of information, and should not be relinquished in favor of some other type of score without good reason.

One of the most obvious suggestions for refining or improving on the raw score is to assign different weights to different items according to their importance. Some improvement in the accuracy and stability of the descriptive information yielded by a test score can nearly always be obtained by differential weighting of the items. However, when large numbers of similar items are involved and where care has been taken to secure a representative sample of items, the need for weighting is very small. In general, weighting of items is very rarely worth while in educational achievement testing. The problem of weighting has been more fully discussed in chapter 10, pages 369–71.

In order to interpret raw scores, they have to be related to many types of information which may affect performance. To facilitate direct interpretation, the scores are converted into derived scores having various characteristics. These vary from the simple percent correct through all sorts of

grade, age, rank, sensed-difference, and maturation scores, to standard scores and scores designed to produce normal distributions or other specified types of distributions in certain groups. Each of these types of scores has certain advantages, and it should be noted that the use of one type does not exclude the use of others. In fact, in most situations several of these types of information are necessary for an adequate interpretation of the test results. The advantages and disadvantages of the various scores are discussed in the paragraphs that follow.

Power scales

Most educational achievement tests are concerned primarily with *power* or level rather than *rate*. There is some theoretical basis (*21*) for contending that the performance on a power test should be expressed not in terms of raw scores, but in terms of the level of difficulty of the most difficult group of items to which the examinee is able to respond correctly some set proportion of the time, such as 50 percent. The test might consist, for example, of ten sets of items grouped according to difficulty, each successive set being more difficult than the set preceding. An individual's score on this test may be 7.2, meaning that he is able to respond correctly to approximately half of the items in the set that has been given a scale value of 7.2, but that he responds correctly to fewer than half of the items in the higher sets and to more than half in the lower. There are a number of very serious practical and theoretical difficulties in constructing such tests, one of the most troublesome being that the items must be highly homogeneous, or highly correlated with one another. Variations in the learning experiences of different pupils are such that few types of items can be expected to be homogeneous, or to maintain a stable comparative difficulty. For this and other reasons, tests reporting performance in terms of level of difficulty are rarely constructed in educational measurement.

Grade equivalents

One of the most widely used methods of interpreting test scores employs the grade equivalent. The grade equivalent of any given test score is the grade level for which that score is the median score for pupils at that level. For example, if a pupil makes a score of 36, and 36 is the median score made by children tested at the seventh month of the fourth grade, he is said to have a grade equivalent of 4.7.

The following step-by-step outline of procedure in setting up grade equivalents may clarify this definition.

1. The test is administered to large samples of pupils in the consecutive grades for which grade equivalents are desired, all pupils being tested

at the same time of the year. Sometimes, if the test is adapted to a single grade or a narrow grade range only, a more advanced form of the test is administered in the higher grades, or a more elementary form is given in the lower grades. In this case, of course, the scores from all of the tests must be expressed in *comparable* terms (see page 699). A distribution of these scores is then made for each grade.

2. The median for each grade is then plotted on cross-section paper against the grade level of each group at the time of testing. One axis of the chart corresponds to the grade scale, the other to the raw score scale.

3. A smooth curve is drawn so as to pass as close to these plotted points as possible. No standards have been developed to indicate how "smooth" this graph should be, and no uniform procedures have been developed for fitting it to the plotted points. Usually, the graph is drawn freehand in accordance with the subjective judgment of the test constructor.

4. This graph is then extended at both ends (extrapolated) according to the judgment of the test constructor. Usually the extrapolation is in accordance with the general curvature of the fitted graph, but no standard or uniform procedures for extrapolation have been developed.

5. Taking each integral unit score successively, a corresponding grade equivalent is read from the graph.

Figure 61 shows a series of such grade equivalent curves as established for the Metropolitan Achievement Test.

It will be noted that the calendar year is considered as being divided into tenths. Some test constructors regard these units as corresponding to months in a ten-month school year running from the first of September to the last of June. This assumes that no growth takes place during the two summer months. Others prefer to regard nine of these units as corresponding to the months in a nine-month school year, and the tenth unit as corresponding to the summer vacation. This assumes that the summer growth in all areas is equal to that in one school month.

Because of their apparent simplicity and ease of understanding, grade equivalents are very popular with teachers, especially at the elementary grade levels. However, there are a number of reasons why they constitute a far-from-ideal type of interpretation at any grade level, particularly for the upper elementary or junior high school grades, and why they are completely impossible for the senior high grades.

1. Grade equivalents assume a smooth curve of growth throughout the *school* year and assume also that either nothing at all happens during the summer months or that growth during that time is equal to the growth in one school month. There is some reason to doubt the accuracy of all of these assumptions. In some functions, such as vocabulary and reading

comprehension, growth is highly related to general mental maturation and probably continues throughout the year more or less independently of specific in-school instruction. In other areas, such as arithmetic or spelling, where learning is more specific, the gains may be small or there may actually be a loss during the out-of-school period. Furthermore, growth in test performance during the in-school period may not be at a uniform rate, but probably is more rapid toward the end of the school year when reviewing is done.

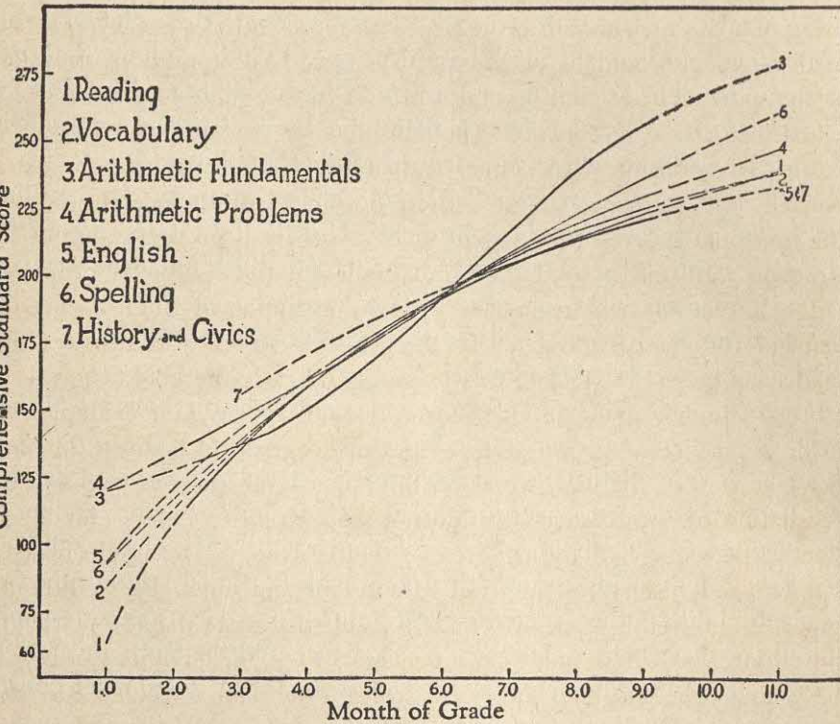


FIG. 61.—Comparison of traditional grade equivalent norm lines for selected subtests of the Metropolitan series. (From *Manual for Interpreting Metropolitan Achievement Tests* [Yonkers, N.Y.: World Book Co.], p. 111.)

2. It is often impossible to establish grade equivalents directly over a sufficiently wide range to describe the performances of all pupils tested. That is, it is usually necessary to arrive at the grade equivalents of high and low scores by extrapolation from actual observations. The proportion of the entire scale that is secured by extrapolation varies greatly from test to test, but for many tests well over half of the scale is secured in this way. These extrapolated grade equivalents at best are informed guesses. There is no established procedure for obtaining them which would insure

that two equally competent technicians would extrapolate a given curve in the same way. Indeed, few technicians could even reproduce accurately their own previous extrapolations, if required to repeat the whole operation without specific memory of the results obtained the first time.

3. Even assuming that the curves are extrapolated in some uniform or reproducible manner, the interpretation of the extrapolated values is most difficult. A raw score in the extrapolated range certainly does not often represent the median score that would actually be obtained if the test were administered to pupils at the corresponding grade level. Possibly these extrapolated scores should be regarded as representing some hypothetical median which would be obtained if the same basic conditions prevailed at the extrapolated grade levels as prevail in the grade levels for which actual medians were obtained. The interpretation of extrapolated values is particularly difficult at the upper end of the grade scale, and the extrapolated values become almost entirely devoid of direct meaning beyond the ninth-grade level. This is partly because there is no continuum of instruction into and beyond the ninth grade for those subjects ordinarily taught in the elementary grades, with the exception of English. Yet the demand for grade equivalents for higher level scores is so insistent that publishers generally yield to this pressure and carry the grade equivalents tables to higher grade levels. World Book Company, for example, has made a practice of giving grade values up to grade 11.0. Even this does not take care of the high scores on the upper-level batteries, much to the frustration of the teachers who do not want to write "11+" for a substantial percentage of the children in their classes. For example, when a standardized history test is given to a bright and especially well-trained group in the eighth grade, it is quite possible that more than 50 percent of the children will have grade equivalents of "11+" or, in other words, will have undistributed grade equivalent scores, *even though the test itself has enough very difficult items to spread out the raw scores of the high-scoring students.*

The meaning of extrapolated grade equivalent scores varies systematically from test to test in different areas due to differences in the degree of overlap of distributions for successive grades. Figures 62 and 63 present graphically the distributions of raw scores on certain tests of reading comprehension and arithmetic for grades six, seven, and eight. It is at once apparent that the overlap in these distributions is much greater for reading than for arithmetic. Below each raw-score scale is the extrapolated grade scale. It is seen that a considerable proportion of the eighth-grade distribution in reading lies beyond the point 10.5 on the extrapolated grade scale, while practically none of the arithmetic scores go beyond this point. The practice

of regarding extrapolated grade equivalents for these tests as *comparable* leads to the inference that very few pupils are as good in arithmetic as are the best pupils in reading, or that exceptional talent in arithmetic is much less common than exceptional talent in reading. However, the

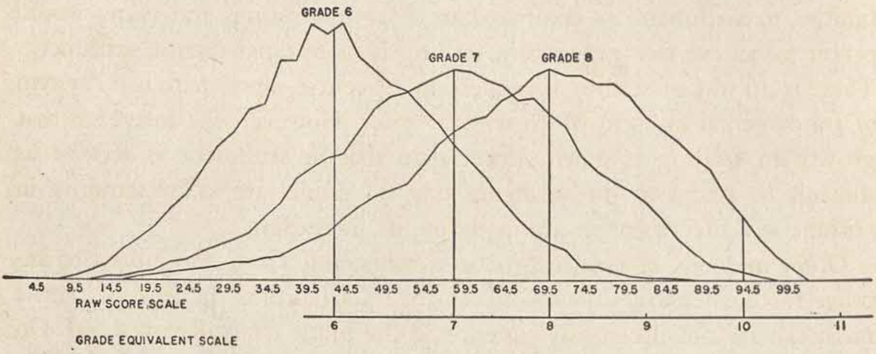


FIG. 62.—Distribution of raw scores on arithmetic test. (Iowa Every-Pupil Tests of Basic Skills, Advanced Form P, for Grades Six, Seven, and Eight. 1944 Program.)

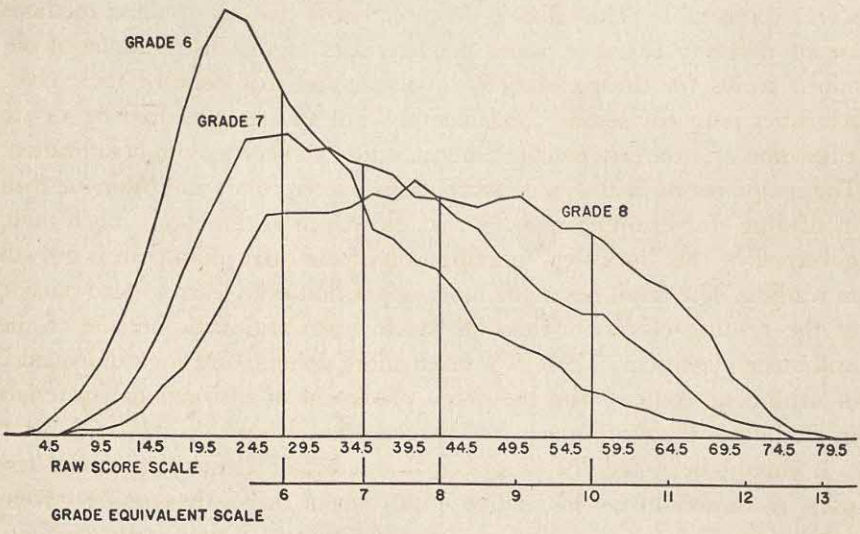


FIG. 63.—Distribution of raw scores on reading comprehension test. (Iowa Every-Pupil Tests of Basic Skills, Advanced Form P, for Grades Six, Seven, and Eight. 1944 Program.)

usual methods of rendering scores *comparable* for different tests (pages 750–60) would start with the assumption that the distribution of talent is exactly the *same* in both areas involved. Thus, the inferences that would be drawn from extrapolated grade equivalent scores are quite inconsistent with those that are drawn from the corresponding raw scores made "comparable" by other methods.

More specifically, the use of a scale that assumes equal within-grade variability from test to test (such as the equi-percentile method described on pages 752–60) would lead to the inference that growth in arithmetic is much more rapid than growth in reading in the elementary school. This inference seems inconsistent with the relatively restricted learning opportunities in arithmetic as compared to those in reading, and many would prefer to believe that progress in reading is more rapid than in arithmetic. There is no way of settling this question, of course, since there is no system of fundamental units to which we may refer. However, the inference that growth in reading is much slower than that in arithmetic is at least as difficult to accept as the inference that no pupils are as outstanding in arithmetic achievement as are many pupils in reading.

✓ Other methods of rendering scores comparable from one subject to another for elementary school achievement tests (such as the equi-percentile method) are not necessarily superior to the grade equivalent method. On the contrary, all things considered, the grade equivalent method is probably the best, at this level, of the methods currently employed for making such scores comparable. The crux of the situation is that all of these methods are of necessity based on some characteristics of the distributions of obtained scores for the populations involved, and that none of these characteristics is in any sense "fundamental," but all are influenced by, or are a function of, arbitrary practices in instruction and curriculum organization. The major reason that grade distributions overlap more in arithmetic than in reading, for example, *may* be that growth in arithmetic is much more governed by the "lock-step" organization of the curriculum than is growth in reading. The pupil has many more opportunities to learn to read outside of the reading classroom than he has to learn arithmetic outside of the arithmetic classroom. There is a much more definite and logical sequence of arithmetic content, and the grade placement of instructional materials in reading is therefore much less definite and uniform than in arithmetic.

It should be noted that a principal interest of elementary school test users is, or should be, in relative improvement or relative progress from year to year in different areas, rather than in status at any particular time. In consideration of this, perhaps the use of a "year's growth" as the basic unit may be as defensible as the use of any other arbitrarily selected unit based on the distributions of obtained scores.

4. The practice of regarding as comparable grade equivalent scores from different tests assumes that the distribution of instructional emphasis for the areas involved remains constant from grade to grade. This is especially contrary to the facts in the case of the informational subjects such as science, social studies, and literature. In some school systems the social

studies are taught as separate geography and history subjects, one being taught one year and the other the next. In some situations science is introduced as a separate subject in the lower junior high school grades while in others it is not taught formally until the ninth grade. These conditions violate the basic assumption of continuity and really make grade equivalents inapplicable for the interpretation of results in these subjects in these grades.

5. All derived scores are subject to some misinterpretation, but because of their directness and apparent simplicity, grade equivalent scores are especially likely to be misused. Manuals regarding their interpretation almost universally emphasize the large amount of overlap in the score distributions of students in consecutive grades. It is nevertheless difficult for teachers and others to realize this and to interpret grade scores properly. For example, if a person entering the fourth grade achieves a grade equivalent score of 5.6, this does not mean that he already knows the materials to be taught in the fourth grade and the first half of the fifth grade. It much more frequently indicates that he has mastered the contents of the first three grades in an unusually thorough fashion. Even the concept of a norm as a mid-score, with half of those in the group falling below it, seems to be grasped with difficulty by those not accustomed to thinking in terms of statistics. Many persons reveal confused thinking about norms by implying that everyone should come up to a norm which has been established as the score achieved or exceeded by only 50 percent of the students in a class group.

6. Many teachers confuse the grade equivalent norm with a standard of work and consider that a given class is doing satisfactory work if the class is up to the norm, without regard to other relevant factors, such as the general intelligence level of the pupils, to say nothing of the commonly found community variations from the national norm due to curriculum differences. This is, of course, not specific to grade scores, but is likely to be the case with many other types of norms also.

The foregoing discussion has been concerned almost entirely with the limitations of grade equivalent scales. As has been noted, all of the other widely used methods of rendering test scores "comparable" are characterized by equally serious limitations at the elementary school level. All that we can do is to employ the method which experience has shown works out best in practice, or that best accomplishes all of the purposes that elementary school achievement tests are intended to serve. To date, the consensus seems to be in favor of the use of grade equivalent scales below grade eight, and of some type of area transformation method in high school and college.

In view of the critical nature of the preceding comments, then, it might be well to conclude this discussion with a summary of the merits of grade equivalent scores.

1. In spite of their many limitations, they probably represent the best of available method for rendering scores "comparable" for elementary school achievement tests. They thus make possible the plotting of profiles of pupil and school achievement, they provide a convenient means of securing weighted composites of scores on different tests, and they facilitate the measurement of *growth* or of development.

2. They do have "direct meaning," always a highly desirable feature in any system of derived scores. The trouble, of course, is that users read into them more or different meanings than they actually possess.

3. They are apparently simple and easy to understand and use. This simplicity is only apparent, of course, but it may result in more effective, more widespread, and more worth-while use of tests and test data by teachers and school administrators than would result from the use of theoretically more adequate but more complex types of scales. They do serve very well, for example, to make teachers more keenly aware of the extremely wide range of individual differences among elementary school children.

Age equivalents

The age equivalent is essentially similar to the grade equivalent, the difference being that age is substituted for grade as the basis for grouping the scores. The age equivalent of any given test score is the *age* for which that score is the median score for pupils at that age. The procedure for obtaining the age equivalents corresponding to the several scores on a test parallels very closely the procedure for finding grade equivalents, but it may be helpful to review the process briefly.

1. The distributions of scores are obtained (on the same test or an articulated series of tests) for several consecutive age groups, regardless of grade placement. An age usually is defined as including those for whom it is the nearest birthday. That is, twelve-year-olds are considered as those whose ages run from eleven years and six months to twelve years and five months inclusive.

2. The median or mean scores of these distributions are plotted against age and a smooth curve is drawn through the plotted points. This curve is extrapolated upward and downward as the needs of the particular situation demand.

3. The age equivalent is read off for each score.

Obtaining the age equivalent has most of the hazards of obtaining the

grade equivalent, plus a few which are peculiar to itself. Let us consider the most important of these hazards.

1. By disregarding grade placement, the age norm makes the tacit assumption that increment in chronological age is the important consideration, not the grade in which recent instruction was received. This factor becomes of less and less importance, of course, as promotion is more and more on the basis of chronological age, *provided* the instruction at grade is well suited to the needs of the individual child.

2. It is very difficult to avoid the effects of selection in setting up age groups. If three grades are tested, three points may be plotted in obtaining *grade equivalents,* but testing three grades will not give even one unselected age group. The test standardization programs at the elementary grade levels, as conducted by the better test publishers, require that the cooperating communities test in at least three grades and that all of the children in these grades be included in the testing. In some instances, where there are substantial numbers of ungraded pupils, even this does not guarantee an unselected population on an age basis.

Regardless of the range of grades tested, the end ages will not be representative and due allowance must be made for the cases not included in these age groups. This is especially important at the upper end of the grade range where the factor of school drop-outs becomes serious and where acceleration takes some smaller proportion of the group out of the grades to go to the secondary school.

3. Age norms prorate the growth over a full twelve-month period regardless of subject, which is an assumption contrary to the one made in setting up grade norms, namely, that no growth or relatively little growth on school subject matter takes place during the summer period. The assumption of steady growth through the year seems more reasonable except that it also goes too far. The truth of the matter probably lies somewhere between the two extremes, with growth in school subject matter during the summer period varying in amount in more or less direct relation to the amount of stimulation to learning existing in the general home and community environment. Age norms obtained by testing pupils in May can be expected to be slightly lower for the same raw score than norms obtained by testing pupils in August.

4. Age norms, like grade norms, are affected by the continuity of instruction over the grade span, by time allotments, curriculum organization, and so forth. However, it seems even less rational to establish age norms for subjects where the instruction definitely is not on a continuing basis, since this means throwing together into *one* distribution cases which have

and have not had instruction in a given subject-matter area. The case of geography versus history is to the point here, of course.

In all of the above discussion, it has been tacitly assumed that the age equivalents are being established along subject-matter lines, rather than for school achievement as a whole. This, unfortunately, is not the traditional course. Traditionally, age norms have been established primarily for some over-all achievement with the primary purpose of comparing achievement with measures of aptitude obtained from an intelligence test. It has been shown that the correlation between two such measures, when corrected for attenuation, is very high indeed—so high that they must be considered as mainly measuring the same thing. This is not surprising when one takes into account the nature of the existing measures. Certainly one need only to look at the current achievement tests to realize that they do depend upon the general response of the organism to the environment as a whole, and not on specific school teaching alone. This is most evident, of course, in subjects such as reading and vocabulary, but it is true of all subjects tested. Some methods of selecting the content also reinforce this observation by putting a premium on items which are passed by larger and larger percents of cases as one goes up the grade (and, incidentally, age) ladder.

On the other hand, our current mental measures are most obviously of such a nature as to be subject to the effects of school learning. Most contain sections on vocabulary, reading, and arithmetic which are quite dependent on school learning.

Age norms are useful primarily for the analysis of the performance of an *individual* and, as such, they are related more to the guidance function of the school than to the administrative. Some of the direct uses of such age values are obvious. For example, the age norm permits a teacher to compare a child's performance in spelling with the performance typical of his age level. This is especially helpful in the case of children coming into a school system from some other part of the country where grade standards may be quite different. However, by all odds the most extensive use of age norms is in the comparison of performance in a subject-matter area (or less justifiably on an omnibus measure of achievement) with other age values, such as chronological age and mental age. The educational age divided by the chronological age and multiplied by 100 yields the educational quotient (EQ). The mental age treated in the same way yields the intelligence quotient (IQ). Dividing the educational age by the mental age and multiplying by 100 provides the accomplishment quotient (AQ). When the educational age is restricted to a single subject-matter area, it

often is called by the subject name, that is, reading age (RA), or arithmetic age (AA).

The AQ technique has been discredited (*5*) very largely as unreliable and unworkable. It has many disadvantages of which the following are perhaps the most important.

1. The AQ is very sensitive to the measurement errors of the measures which enter into its computation. The resulting sampling error is larger because two sampling errors are involved.

2. The AQ is seriously affected by differences in the quality of the groups used in the standardization of the tests from which the age equivalent scores are obtained.

3. The different growth curves which may be found for the intelligence measure used to establish mental age versus the achievement measure used to establish educational age are a source of difficulty, especially in the ranges where extrapolated values are used and unselected age groups are hard to obtain.

4. Several other factors such as the high correlation between educational and mental age equivalents, the disturbing effects of age at school entrance, and the common elements in general mental development as compared with development in school subjects decrease the practical value of the AQ method.

Modal age norms

Some of the disadvantages of age equivalents and grade equivalents are partially overcome, at least, by the use of another type of norm which is based upon a population of children "at grade for their age," that is, by leaving out retarded and accelerated children who are above or below the normal age for their grade. This has the effect of producing a more precisely defined group with which comparisons can be made. The actual modal age group will vary from one community to another, depending upon local administrative policies with respect to entrance age and promotion. However, it would be practically impossible to derive a national modal age norm group by locating and identifying local modal age groups and then combining these in some way. The technique developed by T. L. Kelley, called "ridge route" norms (*12*) was to take that range of twelve months of age in the distribution of ages for the total population at any grade level which showed the greatest concentration of cases, and to consider this the modal age group for that grade.

The effect of basing grade norms upon such modal age groups is to free them of the influence of retardation and acceleration and to make for a "stiffer" standard of achievement. The objection often is raised that

such norms are unfair to the child because he is a part of the *total* grade group and should be compared with the total rather than a selected group. This objection largely misses the point that the more precisely defined and homogeneous the comparison population is, the more meaningful becomes the comparison made with it.

It is definitely true, however, that such a modal age population is slightly above average in intelligence, since it includes only individuals who started at the usual age and were regularly promoted. In contrast, at most grade levels, the traditional grade group is below average in intelligence because it includes a number of older students who have been held back.

Some of the more important characteristics of the modal age group other than those mentioned above are as follows:

1. The modal age group for each successive grade level is usually one year older than for the previous grade.

2. The modal age group as herein defined in any substantial population under existing conditions has been found to contain about half the cases in the grade.

3. Although IQ range is reduced, the variation of IQ's within the modal age range is still very great. Table 11 shows the range of IQ's in the modal age group at each grade level in the Metropolitan National Standardization program. The numbers of cases in this table are the numbers in the sample analyzed and constitute a random sample of about 20 percent of the total number of children tested at each grade level. This population also was restricted to white children in public schools in order to achieve an even more precisely defined group.

Centile ranks and percentiles

Considerable confusion exists in practice concerning the meaning of "percentiles," "percentile ranks," "centile ranks," and "percentile scores." Sometimes reference is made to the percent of the given group earning scores *lower* than a given integral score, sometimes to the percent scoring *at or below* the given integral score, and sometimes to the percent *below the mid-point* of the integral score interval. Sometimes the reference is to the raw score below which a given percent of the distribution lies, and sometimes to this percent itself, etc.

For the purposes of this discussion, the "centile rank" of a given integral score in a given distribution of such scores will be defined as the percent, rounded to the next higher integral value, of the scores in the distribution which lie below the mid-point of the integral score interval, including half of the scores within the given interval. An individual's centile rank may thus be regarded as his rank in a standard group of 100 indi-

TABLE 11

Distribution of Pintner IQ's for Modal Age Group
Metropolitan National Standardization

IQ	Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8		Grade 9	
	f	%ile	f	%ile	f	%ile	f	%ile	f	%ile	f	%ile	f	%ile	f	%ile
160							1	100.0			1	100.0	2	100.0	1	100.0
150–159	2	100.0	1	100.0	4	100.0	9	99.98	7	100.0	1	99.98	1	99.94	1	99.94
140–149	29	99.96	6	99.98	11	99.9	44	99.8	37	99.9	8	99.96	6	99.91	4	99.9
130–139	125	99.4	93	99.9	130	99.8	212	99.2	170	99.3	98	99.8	40	99.7	12	99.7
120–129	387	96.9	508	98.4	662	98.0	608	96.1	537	96.7	389	97.6	244	98.5	108	99.0
110–119	1,127	89.3	1,333	90.2	1,503	88.9	1,392	87.3	1,273	88.4	922	89.1	683	91.0	433	93.0
100–109	1,734	67.1	1,897	68.8	1,953	68.2	2,193	67.2	1,957	68.7	1,426	68.8	920	70.0	540	68.9
90– 99	914	33.0	1,537	38.3	1,712	41.4	1,516	35.4	1,575	38.4	1,094	37.4	837	41.7	426	38.9
80– 89	573	15.0	674	13.6	971	17.8	747	13.5	732	14.1	462	13.4	422	16.0	207	15.2
70– 79	160	3.7	151	2.8	287	4.5	141	2.6	158	2.8	122	3.2	87	3.0	62	3.7
60– 69	24	.57	19	.37	36	.55	40	.61	18	.34	24	.53	10	.34	4	.28
50– 59	5	.1	3	.06	4	.05	2	.03	4	.06			1	.03	1	.06
40– 49			1	.02												
Number	*5,080*		*6,223*		*7,273*		*6,905*		*6,468*		*4,547*		*3,253*		*1,799*	
Median	*103.6*		*103.5*		*102.8*		*104.5*		*103.5*		*103.5*		*102.0*		*102.5*	

viduals of the specified type. This definition means, of course, that centile ranks will run from 1 to 100, rather than from 0 to 99.

A given "percentile," on the other hand, will be defined as the point on the scale below which a given percent of the distribution lies. This implies *interpolation* within a unit interval. Theoretically, this interpolation should be with reference to the *smooth curve* that best describes the distribution. Actually, percentiles are nearly always secured in practice by linear interpolation which assumes a uniform or flat distribution within the interval. The limits of the interval are usually taken as .5 unit above and below the integral value.

The point in question may be identified either in terms of its position on the score scale, or in terms of the percent of the distribution involved. It seems best to let *percentile* refer to the position on the scale, and *percentile score* refer to the percent. Thus, a score of 37.4 may be the 19th percentile point, and the percentile score corresponding to a score of 37.0 may be 18.6; that is, this score is above 18.6 percent of those in the group.

It is clearly desirable for test technicians in general to agree upon an exact, standard terminology. The definitions suggested above probably come as close as any to the consensus of current practice, and it is hoped that these definitions will secure more widespread acceptance. The practice of designating as "centile ranks" either the percent *below*, or the percent *at and below* a given integral value, should be discouraged.

To illustrate the computation of centile ranks and percentiles, let us suppose that 40.2 percent of the scores in a particular group are 23 or below and 38.2 percent are 22 or below, 2 percent being precisely 23. To compute the 40th percentile of the score distribution, the proportion of the 23rd score interval below which 40 percent of the group falls is estimated by linear interpolation $\frac{1.8}{2.0} \times 1.0 = .9$. This value is added to the lower limit of the score interval, taken as 22.5, and the 40th percentile is reported to be equivalent to a raw score of 23.4. This does not indicate the centile rank to be assigned to a score of 23 or any other actual score, however. An individual obtaining a score of 23 is considered to have exceeded half of the individuals in the group who also obtained that score, including half of himself if his score is in this distribution. His score, therefore, exceeds those of 39.2 percent of the group. Since his performance is better than slightly more than 39 of those in a group of 100, his performance entitles him to the 40th rank in this group and his centile rank score is said to be 40.

Certainly there are few more readily grasped methods of presenting rela-

tive standing with respect to all types of groups than the centile rank, and it is usually desirable to obtain not one but several centile ranks in interpreting an individual's score. For example, his score might be compared not only with the distribution of those in his own class, but also his rank in his own school and his own city are of interest, in addition to his standing on a national basis. His rank with respect to those in particular types of schools and his rank with respect to applicants for scholarships or college entrance are frequently desired. Finally, it is of great value to compare his score with those of individuals successfully engaged in various types of civic and occupational activities.

Centile ranks, however, do not seem very desirable as basic units to use in statistical analysis. It does not appear either reasonable or useful to regard rank scores as representing equal increments of ability throughout the scale. Ordinarily the score distance is much greater in terms of ability between the 5th and 10th or the 90th and 95th percentiles than between the 45th and 50th or the 50th and 55th.

Percentile norms

A percentile norm is an estimate of a population percentile. As such, it must be distinguished from the observed value of a percentile computed for a particular sample (particularly if computed by linear interpolation within an interval). The estimate should be as little influenced as possible by purely chance variations or sampling errors in the frequencies in individual intervals in the sample distribution, and should also be free from variations associated with the size of the interval resulting from the assumption of a rectangular distribution in the interval. This means that a percentile norm is best estimated from a sample if it is read graphically from a smoothed ogive, rather than computed arithmetically from the observed interval frequency on the assumption of such a rectangular distribution. Most important, this method should result in considerably more stable estimates, since any value read from a properly smoothed curve depends on the frequencies of a number of intervals rather than upon only one. This is the method that is frequently employed in actual practice, and the foregoing is thus only a way of making the definition of percentile norm consistent with the practice. Incidentally, if the distribution of raw scores is roughly normal in form, the best results in smoothing will probably be obtained if the ogive is plotted on arithmetic probability paper. The reasons for this are given on page 728.

There has been some confusion in practice concerning percentile norms, since they depend on percents and not on integral percentile values. It can be readily seen that at the points on the score distribution where 1 percent

of the cases are spread out over several score points, it is quite possible that 83.4 percent of the cases will fall below the mid-point of the score interval 64, but that the 84th percentile point will be closer to this mid-point than will the 83rd percentile point. This would be found if there were relatively few scores of 62, 63, and 64, and many of 65.

Sensed-difference scores

Various methods (*14*) of obtaining units by using comparative judgments have been proposed. These are usually based on just-noticeable-differences or equally-often-noted-differences. Handwriting scales have been developed using this principle. When applied to more typical educational test content, such judgments require unusually complete and unbiased knowledge regarding the information and proficiencies involved in the materials being compared. However, it is very difficult to find even a moderate number of judges who are sufficiently informed to compare the relative achievements of a significant group of subjects. If judgments are to be based on previously accumulated information, many types of extraneous factors are introduced. For example, few people who have previously seen quantitative data regarding the individuals involved can prevent those data from playing a large part in their judgments even though they are of only very slight relevance to the immediate situation. Knowing that a person had a high score on a computational test would almost invariably raise a judge's estimate of his word knowledge, even though the two abilities involved were quite independent. The extent and type of this information will vary a great deal from one situation to another. Similarly, personal attitudes and impressions regarding irrelevant aspects of the individual tend to influence judgments of specific abilities.

It is quite possible that discriminations would be finer at a level having administrative importance. The judgments with respect to passing or failing may frequently be more accurately made than judgments at other levels. It seems unreasonable to expect that significant behaviors will have been equally well observed at all levels. Therefore, the units are likely to be too dependent on the experiences of the judges.

It would be possible to have judges estimate the difficulties of items directly. This would be open to the same objections regarding dependence on the experience of the judges as mentioned above. The judges' estimates must be based on knowledge of the experiences of the individuals who are to take the items, as well as information regarding the learning difficulty of the materials themselves. It does not appear that the type of sensed-difference scores based on comparative judgments of loudness, pitch, or brightness have any very direct counterpart in educational measurement.

Learning curve scores

An attempted refinement on age, grade, and years-of-study scores has been proposed by assuming a general form for the growth or learning curve in all functions. The most widely publicized proposal along this line is for the use of Courtis "isochron" scores (*4*). The derivation of these scores assumes that all growth and learning are basically in accord with the Gompertz growth curve. In terms of isochron units 0 means no learning at all and 100 indicates complete maturity. Isochron scores indicate directly the percent of complete maturity attained in terms of time units. Each isochron unit is one one-hundredth of the period of time from the beginning of development to complete maturity, and the raw score is converted to an isochron score. A score of 56 would indicate that the results achieved are those to be expected after spending 56 percent of the time necessary for complete learning.

These units appear to offer simplicity and comparability. The principal disadvantage is the complexity of the functions to be measured in education. The objectives of a subject such as geometry are so varied that it would take many series of curves to provide an adequate representation of progress along each of these lines in this subject. On the other hand it does not seem reasonable to prepare one single curve to depict progress in the whole field of geometry. Another very serious practical limitation is the difficulty in identifying the upper limit to be used in defining complete maturity. There are very few areas in which anyone ever attains complete maturity. It is also extremely difficult to place the zero point in a practical situation. Therefore, although learning expressed in equal time units appears at first consideration to be an excellent idea, it does not seem practical for the typical kinds of educational measurement in current use.

Standard scores

A useful type of transformation which will provide any desired mean and standard deviation for the derived scores can be obtained by adding or subtracting a constant value to or from all raw scores and multiplying the results by another constant. Such a linear transformation differs from most of the types of derived scores previously discussed in that all differences between individuals retain their same relative values. "Standard measures" or "z-scores" use a pair of constants which result in a mean of zero and a standard deviation of one for the group used as a standard; that is, the mean raw score is subtracted from each score and then each difference is divided by the raw-score standard deviation. Other variations of this technique use mean values such as 50, 100, or 500, with standard deviations of

10, 20, and 100 respectively. Such scores simplify interpretation and increase comparability. However, two forms of a test sometimes have a curvilinear relation between their scores. This results in non-comparable standard scores from the two forms. In many instances there are other reasons to suspect the uniformity of the units in the raw-score scale. If the raw-score scale is distorted, with large units at the low score end due to too few easy items and small units at the high score end because of many very difficult items, the standard scores will have precisely the same defects. The standard scores described above should be differentiated from various other types of transformed scores such as those described in the next section which are sometimes also called standard scores.

The shortcomings of the linear standard scores suggest the need for some other type of unit. The fact that many distributions in raw-score units tend to have a bell-shaped form suggested that perhaps units giving a specified shape to the distribution might be stable and useful.

Scores normalizing a single distribution

One of the earliest suggestions for scaling a test was to modify the units of a distribution of scores so that the distribution in terms of the new units follows the normal curve. There is considerable evidence favoring the belief that such units may be more basic and less subject to the decisions and procedures of the test constructor than raw scores. One point is that distributions of raw scores for groups at various levels do tend to approximate the normal curve. Since test constructors tend to provide a range of items covering the scale fairly evenly, this finding seems significant. That raw-score distributions deviate from normality in an erratic and unpredictable manner and in all possible directions lends further support to the belief that these are chance fluctuations about a stable form.

Another type of evidence favoring the normal distribution as a basic underlying form comes from analogies with physical measurements and with measures of complex reaction times and motor skills, which tend to be of the normal form for populations homogeneous in related measures.

Many investigators have used score transformations which normalized their distributions. One of the best-known types of scores of this type is the T-score (*15*). These have been defined by McCall as scores with a mean of 50 and a standard deviation of 10 providing a normal distribution for unselected groups of twelve-year-olds. One of the problems with this type of score is selecting the distribution for which the scores are to be normal. Obviously groups of scores having different means or different standard deviations cannot be added to or taken from the total distribution without affecting its shape. Another problem is that unless the group is extremely

large, score units at the high and low ends of the scale will be rather unreliably scaled. These and other difficulties in the establishment and use of normalized scores will be considered later (pages 727–41).

Scores normalizing a series of overlapping distributions

The problems just raised concerning the normalizing of the scores for a single distribution can sometimes be overcome by a technique that attempts simultaneously to normalize each of a number of overlapping distributions. This technique should result in a more accurate scaling at the extremes of the range, since the use of a series of overlapping groups provides a better coverage of the complete range than can be expected from a single group involving the same number of cases. This scaling technique is described in detail in a later section (pages 732–39).

One of the most interesting and convincing features of this type of scaling is the tendency for the units to be invariant for scalings based on groups with quite different means and standard deviations. This phenomenon is most strikingly illustrated by translating the raw scores (group A) for a large group to centile ranks obtained from another distribution (group B) for which the mean and standard deviation are quite different from those in group B. The distribution of these centile ranks for group A will almost invariably depart markedly from the rectangular form of the original centile rank distribution as obtained for group B. On the other hand it is usually found that if the raw scores of group A are translated into normalized scaled scores obtained from the distribution of group B, these scores will also tend to be normally distributed. This is the most crucial test for any system of scores: *Will new groups of a type homogeneous with the original group but having smaller or larger means and standard deviations than the original group tend to have score distributions similar to that assumed for the original group or groups on which the derived scores are based?*

There appears to be considerable evidence favoring the assumption of normality, either with single or overlapping distributions. In practical work, a point needing verification is the precise definition of the groups to be used in the scaling. In deriving the Scaled Scores (6) of the Cooperative series of tests first introduced in 1937, the decision was made to obtain units which tend to make the distribution of the test scores obtained by the students in a single grade studying a particular subject in the same school system approximately normal. It was hoped that the procedures used would tend to do this for all school systems. It was believed that any attempt to select a larger unit such as a state, region, or the entire nation would some-

times lead to less tenable assumptions. For example, if a state is made up of two large metropolitan areas containing half the population and the remainder is in small rural communities, it does not seem reasonable to assume normality in each metropolitan area and in the rural schools and at the same time suggest that the entire group in that grade or subject for the state be regarded as normally distributed. Although certainly each situation must be considered on its merits, there seems to be much to be said for using groups in which the course of study, general quality of instruction, textbooks and other instructional materials, and the length of the school year have been fairly homogeneous for all individuals in the group.

Scores from overlapping distributions involving the assumption of other types of curves

Even though the distributions of scores are confined to students with as homogeneous a background as suggested above, it still must be recognized that the assumption of normality is very likely to be an oversimplification in the practical testing situation. In the first place, unless the trait being measured is a pure trait, a change in the level of development of one of its components will alter the form of the distribution. Secondly, it should be noted that a skewing of the curve, with a long, drawn-out tail at the upper end, may be caused by the operation of certain especially favorable factors such as hobbies, clubs, trips, and special home conditions. These factors sometimes cause certain individuals to attain levels of educational development far beyond the range usually associated with their scholastic status. As a third point, it may be mentioned that the administrative policies of the school system involved in the scaling, such as requiring students to repeat grades, may cause a truncation of the lower part of the distribution of scores or at least reduce the number of these cases by eliminating the least able students. The number of high scores in a grade may also be reduced somewhat by the acceleration in school of the more able students.

These considerations lead to a question as to whether a skewed curve might not provide a better fit for the obtained data. To allow this possible skewness to appear, E. F. Gardner, in developing his K-units (*8*), assumed a Pearson Type III distribution, and sought to obtain units which would distribute scores from overlapping grade groups in accordance with curves of this type and keep the same proportion of the fitted curve for each of the overlapping groups below each of several points on the distribution as was the case with the raw-score units. In his scaling, in addition to establishing the ratio of the standard deviations of the overlapping groups, he determined the appropriate skewness value for each group. Since a Pearson

Type III curve with no skewness is a normal curve, this procedure allows one more parameter to vary than does the scaling using normal curves.

A test of any scaling procedure is the extent to which the scores from a homogeneous group not a part of the scaling group tend to distribute themselves in accordance with the type of curve assumed in the scaling procedure. This test should always be applied to new data at the conclusion of the scaling operations.

Probability of success norms

An additional type of data for use in reporting and interpreting test scores is a value indicating the likelihood of achieving a specified degree of success in a certain activity which has been determined on the basis of follow-ups of groups of persons obtaining this score under similar conditions. Thus, it is of value to know that a person receiving a raw score of 76 on a stenographic test is in a group 92 percent of whom performed in a satisfactory manner with respect to taking and transcribing dictation when subsequently assigned to secretarial positions in specified types of companies. Such information is very specific, and scores based on this type of information would have limited general value. In the specific situation to which it applies, it is the ideal type of information. Such norms are obtained by following up a sample and determining the relative frequency of success for each score group within the sample.

Technical Problems in Establishing Scores and Norms

Basic units

It has been pointed out in the preceding sections that raw-score units do not generally have the characteristics of a desirable system of units. An interval scale with equal units throughout the scale is necessary to obtain valid use of scores in the calculations necessary to obtain measures of growth, trait differences, means, standard deviations, and product-moment correlation coefficients. In obtaining basic units, it has been customary to accept the raw scores as indicative of the proper rank order of the individuals in a group. These raw-score units are then transformed to give the derived units certain new properties considered desirable. Such conversions can do nothing to overcome the defects in the raw scores resulting from improper weighting of components, poor selection of content, and inferior types of items. These are problems of validity regarding the purity, relevance to the real situation, representativeness of the universe of items defined by the outline and specifications for the test, and predictive efficiency in a specified situation. The discussion in this section will concern itself only with

methods for deriving basic units which preserve the rank order established by the raw scores.

Linear transformations

The most widely used scaling procedure is the simple linear transformation. Linear transformations consist in effect of sliding the score scale up or down on the distribution and expanding or contracting the units in a uniform manner. This is done by setting up standard scores with any specified mean and standard deviation using a procedure of the type

$$X' = aX + b = \frac{\sigma'}{\sigma}X + \left(M' - \frac{\sigma'}{\sigma}M\right),$$

where X' is the new standard score, X is the original raw score, and a and b are constants. The first constant, a, is the ratio of the new and original standard deviations. The second constant, b, is the difference between the new mean and the product of the ratio of the standard deviations and the old mean. This transformation does not affect the shape of the distribution in any way. If a score is five standard deviations from the mean in the original units, it will still be exactly five standard deviations from the new mean in the new units.

Area transformations

In contrast with the preceding, area transformations, which are nonlinear transformations, change the units in such a way that the proportion of the scores falling in a particular interval will conform precisely to the proportion of the area under that section of a curve of the specified form. Suppose, for example, that it is desired to use a nine-point scale along which the unit is equal to one-half of a standard deviation, and in which the score 5 represents the median. To transform to this single-digit system of scores, called "stanine scores," the lowest 4 percent of the scores would be given a value of 1; the next 7 percent, 2; the next 12 percent, 3; the next 17 percent, 4; the next 20 percent, 5; the next 17 percent, 6; the next 12 percent, 7; the next 7 percent, 8; and the highest 4 percent, 9. These are the proportions (rounded to nearest whole values) which are found in the corresponding $\frac{1}{2}\,\sigma$ intervals in a normal distribution. For instance, a table of area relationships under the normal curve will show that 9.9 percent of the area lies between the median and a point one-quarter of a standard deviation away, hence $2 \times 9.9 = 19.8$ percent, or 20 percent rounded, lies in the median interval.

The percent to be assigned scores of 6 is similarly determined by noting

from the tables that the area from the mean to three-quarters of a standard deviation above the mean contains 27.3 percent of the area under the normal curve. Subtracting 9.9 percent from this, it is found that the next 17 percent of the scores above the median after assigning the first 10 percent scores of 5 should be assigned scores of 6. The other percents are determined similarly.

Since this "normalizing" method of scaling is one of the most widely employed in educational measurement, it may be well to provide a specific example illustrating the computational procedure. Perhaps the best way of normalizing a distribution of test scores is that which employs arithmetic probability paper. This is cross-section paper in which the horizontal lines are spaced as are the percentile points in a normal distribution, while the vertical lines are uniformly spaced. When the relative cumulative frequencies for a normal distribution are plotted on this paper, the graph is a straight line, whereas on ordinary paper it is the familiar S-shaped ogive.

Table 12 presents the frequency distribution of raw scores on a vocabulary test for a sample of 515 seventh-grade pupils. Figure 64 shows the cumulative frequency graph for these data plotted on arithmetic probability paper. In this case it is desired that the mean of the derived scores be 15, and that the unit be one-fifth of a standard deviation. The derived scale is laid off along the left-hand margin. The unit is obtained simply by dividing the distance from the median to the 84.13 percentile (which is 1 σ from the mean) into five equal parts, the value 15 is arbitrarily assigned to the median, and the rest of the scale is laid off from the median. The derived score corresponding to any raw score is then obtained by erecting a perpendicular from the desired point on the raw-score scale until it intersects the graph, and then reading across horizontally from this point of intersection to the derived score scale. Thus, the derived score corresponding to a raw score of 20 is 11.3.

The facts that the graph is a straight line for any normal distribution of raw scores, and that it shows relatively little curvature for distributions that do not depart markedly from the normal, make it relatively easy to smooth out chance irregularities and to extrapolate to outlying values. The smoothed graph (solid line) in Figure 64 was drawn with the aid of a French curve to provide a close fit to the plotted points. The reading of high and low values can be done much more satisfactorily when arithmetic probability paper is used than when ordinary cross-section paper is employed. Finally, the use of arithmetic probability paper obviates the need of referring to a table of area relationships under the normal curve, and is thus a very decided timesaver.

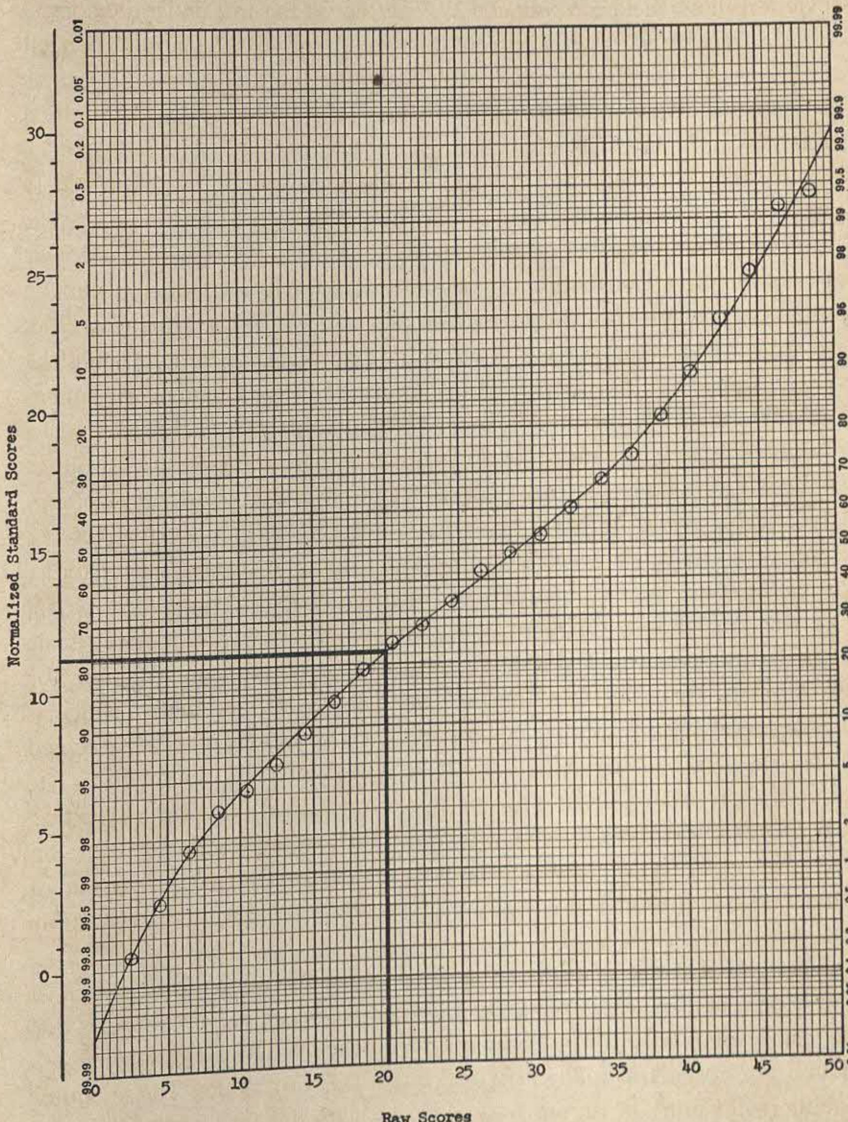


FIG. 64.—Illustrative use of arithmetic probability paper in normalizing distribution of scores in Table 12.

Area transformations are not limited to transformations to the normal distribution. It is possible to transform to any type of distribution whatsoever. The operations are, of course, greatly simplified if values for the areas under the curve corresponding to specified deviations from the median value are available in tabled form, or if paper similar to arithmetic probability paper is available. Tables of the normal distribution are readily available in various forms. For Pearson's Type III curve, Salvosa (*19*) has prepared a useful set of tables.

TABLE 12

Distribution of Scores on a Vocabulary Test for 515 Seventh-Grade Pupils

Scores	Frequency	Relative Cumulative Frequency
49–50	3	100.0
47–48	1	99.4
45–46	9	99.2
43–44	15	97.5
41–42	30	94.6
39–40	35	88.7
37–38	43	81.9
35–36	31	73.6
33–34	39	67.6
31–32	39	60.0
29–30	26	52.4
27–28	25	47.4
25–26	43	42.5
23–24	30	34.2
21–22	20	28.3
19–20	28	24.5
17–18	29	19.0
15–16	22	13.4
13–14	15	9.1
11–12	10	6.2
9–10	6	4.3
7– 8	8	3.1
5– 6	5	1.6
3– 4	2	.6
1– 2	1	.2

Selection of basic group

One of the most important decisions with respect to any method of scaling based on form of distribution is in regard to the selection of the group or groups which presumably possess this specified form of distribution.

A possible method of scaling a school achievement test is to normalize

a single distribution representing a *very large* and relatively homogeneous group of *schools*. The scales for the Iowa Tests of Educational Development, for example, were established by normalizing a distribution of approximately 10,000 scores, obtained by taking a *representative* sample from pupils tested in 290 Iowa high schools (each school being proportionately represented). The assumption was made that the distribution for this sample was "fundamentally normal."

There are many instances in educational achievement test construction, however, in which the test constructor cannot employ a sample representative of a large number of schools. In many instances, the problem is one of doing the best that is possible with the data from a relatively small number of schools. In this case, the procedure of normalizing several overlapping distributions seems theoretically the best to employ. Under this procedure, a number of factors should be carefully considered before defining the characteristics of the groups to be used. Some of these are listed below:

1. *Amount of instruction.* If it is assumed that a group having had ten months of instruction in a subject distribute themselves in accordance with some fundamental form of distribution curve, it becomes ridiculous to suppose that a second group of unspecified size having had only seven months of instruction can be added to the original distribution without expecting a change in the form of the combined distribution. It does not appear that school systems and classes in which either the time in months devoted to the subject, the proportion of the school day, or the frequency of class meetings during the week are greatly at variance can be combined in unspecified proportions without disturbing the basic form of distribution (unless a very large number of systems is involved, and these school-to-school variations may themselves be assumed to be normally distributed).

2. *Quality of instruction.* It does not seem reasonable to combine distributions obtained from two groups of students when the teachers' salaries in one group are twice as high as those in the other. It seems unreasonable to believe that both school systems are getting the same quality of teaching under these circumstances. Perhaps a better index of quality of teaching would be scores of teachers on the National Teacher Examination, or the amounts of their education. In any case, certainly the quality of instruction can be expected to produce differences in achievement with respect to educational objectives.

3. *Content of instruction.* Perhaps the largest factor in determining the content of instruction is the textbook. The extent to which reference books are used and, more broadly, the course of study followed are also important variables. Certainly there are sufficient differences in texts and courses of

study to suggest important differences in educational outcomes if one is used rather than another. The combining of a small number of samples having such diversity in unknown proportions certainly would not seem to contribute to producing a stable form in the combined distribution.

4. *Character of previous instruction.* Another variable which should be considered is previous instruction. It does not seem reasonable to combine scores from two groups of intermediate algebra students if one group had previously studied only one year of elementary algebra and the other had had three years of mathematics prior to beginning the course. In the elementary grades similar problems of grade placement of subjects occur which seem likely to have significant effects on the outcomes of instruction.

5. *Aptitudes for instruction.* Perhaps the most important single consideration is the aptitude of the individuals in the group for this particular type of instruction. To expect a stable form of distribution, the selection of individuals should have been based on a complex of natural causes, not a few administrative decisions. Certainly groups from school systems should include all of the students enrolled in the particular course or a random or representative sample of them. Situations in which a sharp cut-off on the basis of special aptitude is employed in selecting individuals for the courses are especially troublesome in trying to establish a stable form of distribution. The combination of groups from school systems in which the general aptitude level varies substantially also introduces difficulties in interpreting the expected form of distribution for the total groups.

For the reasons discussed above, the groups selected for scaling in establishing the Scaled Scores for the Cooperative tests were from school systems each of which was under the administrative supervision of a single board or individual. It was stipulated that all students enrolled in the particular course should have been tested and their scores included in the distribution used.

Normalizing overlapping distributions

In order to cover adequately the range of possible raw scores and to obtain large enough samples to give stability to the scaled units, it is sometimes desirable to attempt to normalize a number of overlapping groups. There seems to be no optimal number of groups. Other things being equal, the larger the number of groups and the larger the number of cases within each group, the more accurate the scaling. It appears, on the other hand, that in most instances of practical scaling of educational achievement tests at least three groups are necessary to cover the range of scores adequately.

It also appears that homogeneous school groups of sufficient size to be worth while can be expected to contain anywhere from a hundred to several thousand cases, with most of them having less than a thousand.

The units of the new score scale are based on the stability of proportions, since they depend on the comparison of the proportions of each of the overlapping groups which lie above and below specific points. The standard error of a proportion is $(pq/N)^{\frac{1}{2}}$. Thus, for 2,500 cases, the determination of the point corresponding to half the group has a standard error of 0.01 in terms of the proportion of cases. This corresponds to 0.025 standard-deviation units at the middle of a normal distribution. Similarly the same number of cases yields a standard error for the one-tenth versus nine-tenths split of 0.006 in terms of proportion of the group. This corresponds to 0.034 standard-deviation units at this point on the normal curve. The standard errors of these points would be five times as great if there were only 100 cases in the groups being studied. In this way the technical worker can decide on the necessary size of groups to attain the degree of precision in establishing units which he believes is appropriate. The number of groups necessary to cover the range of scores will depend on the reliability and internal consistency of the items in the test. In practice it can easily be seen whether the range of scores is adequately covered. For theoretical purposes the considerations in the following two paragraphs provide an indication and a guide.

The standard deviation of a test in terms of raw-score units is largely a function of the intercorrelations among the items, provided the difficulty level is appropriate. For a 100-item test having close to zero reliability, the standard deviation of the scores, if all items are answered correctly by 50 percent of the group, will be $\sqrt{Npq}$ or 5.0. The range of raw scores would be about 25 points in a group of 100 cases or 35 points in a group of 2,500 cases. The fact that the best estimate of the standard deviation as determined by the proportion of cases between two points is obtained when these points are close to the 7th and 93rd percentiles in a normal distribution (*10*) suggests that the means of the successive distributions should be not more than one and one-half standard deviations apart to provide a stable amount of overlap. For a normal distribution 7 percent of the cases fall beyond the point one and one-half standard deviations above the mean. Thus, for groups taking the 100-item test described, in order to obtain a fairly satisfactory estimate of the relative variabilities of the two groups on the basis of their amount of overlap, it is proposed that if the mean of the lowest group is at a score of 10, the mean of the next group should be around 17.5, the next about 25.0, the next near 32.5, and so

forth up to at least a score of 90. This suggests that in this instance between 10 and 15 groups would be necessary to cover the range of raw scores from 0 to 100.

Suppose the test items have sufficient internal consistency so that 90 percent of the variance in the total test scores is due to the covariances. This would be true if the items correlated to the extent of about .10 with each other, for in that case the variance of the scores would be equal to the sum of the 100-item variances, 25, plus the sum of the 9,900 cross-product terms each of which would be .10 as large as the item variance terms. The sum of the cross-products terms would in this case be 9.9 times 25, and the covariance would contribute 247.5 of a total of 272.5 points to the variance. This is just a little more than 90 percent of the variance. The standard deviation is about 16.5 score points, and the range for groups of 100 and 2,500 cases is respectively about 82 and 100 (the maximum possible). In this case, about three groups would be necessary to cover the raw-score range. The first group might have a mean of 25, the second of 50, and the third of 75.

Having selected the groups and tabled the frequency distributions in terms of raw scores, the next problem is to obtain estimates of the positions of the means and the relative sizes of their standard deviations in terms of the basic units being derived. If curves of Pearson's Type III are being investigated, it is also necessary to establish the appropriate skewness for each distribution. Various procedures for performing these operations have been developed.

The first step is always based on a comparison of the distance between the same two selected points on the raw-score scale in terms of standard-deviation units of each of the distributions transformed into the assumed form, for example, normalized or with a specified degree of skewness. This is done in the case of the assumption of normality by computing the proportion of cases below each of the points for the first group and looking up the values for the points corresponding to these areas in a table of the normal probability function. The difference between the two tabled values provides an estimate of the distance between the two points in terms of standard-deviation units of the first group. The process of looking up values corresponding to the areas below the two points is repeated for the second group, and the difference between these two values is the distance between the two points in terms of standard-deviation units for the second group. The division of these two values provides the required ratio of the standard deviations for the two groups. This is illustrated in Figure 65.

If estimates of this ratio are to be based on only one pair of points, it is

desirable to have them as far apart as possible so as to reduce the effect on the ratio of sampling errors in determining the distances for the two points. It can be seen that increasing the distance tends to reduce the effect of a fairly constant error in fixing the point on the basis of frequencies. On the other hand, if one of the points on which the distance is based is far out on the tail of the distribution and, therefore, is determined by only a few cases, its sampling error will increase so much that the advantage of the greater distance will be more than lost. Probably sufficient cases in the tails of the distributions overlapping the medians of adjoining distributions so that between 5 and 10 percent of the scores are involved will be found

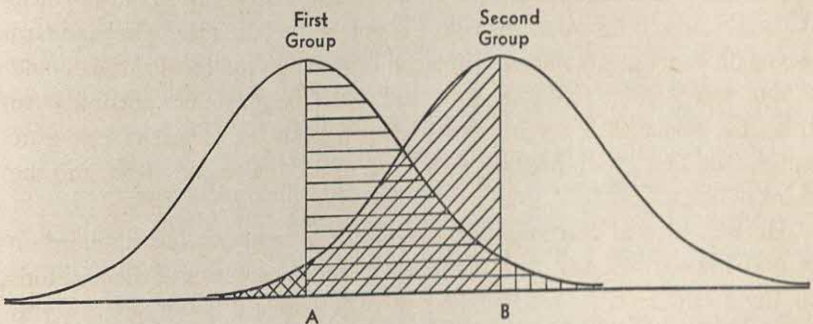


FIG. 65.—Normalized curves for two groups

optimal, as indicated in the previous discussion on the necessary number of overlapping groups. Estimates of the distance between the medians can be obtained from the proportions of the two curves falling below only a single point if the two standard deviations are known. The distance between the medians of the two curves is equal to the difference between the products of each of the two deviation values for the point and its respective standard deviation in terms of the basic units of the scale. The standard deviations must both be expressed in terms of the basic units which are fixed in terms of the standard deviation of one of the overlapping groups which has been arbitrarily designated as the basic group.

$$M_{A_1} - M_{B_2} = Z_{1_A}\sigma_1 - Z_{2_A}\sigma_2$$

There is no good reason for limiting the estimate of the ratio of the standard deviations to the values based on only two points, nor for limiting the estimate of the distance between means to data from a single point. The most accurate method for deriving first estimates of the ratios of standard deviations and the difference between means is to obtain the best fit (least squared error) straight line to a large number of points with each

point weighted inversely as its variance error. A good approximation to this can usually be obtained by plotting several points each corresponding to the two deviations from the means of the two groups being compared in terms of a unit normal curve. A straight line can be fitted to this graphically. This procedure is especially appropriate where these are to be used as first approximations in an iterative procedure. Plotting a point corresponding to each score interval for which the two distributions overlap would be the most accurate procedure. Satisfactory first approximations can usually be obtained by using five or six well-chosen points. This procedure has the advantage of indicating graphically whether a straight line will provide a good fit to the data. If it will not, then the form of distribution assumed cannot be regarded as in accordance with the obtained data.

It is not necessary to plot the points to obtain a combined estimate of the desired values. An average or weighted average value for the ratio of the standard deviations may be obtained by calculating values as described above for two points and repeating for all of the independent comparisons between several other points. To obtain a first approximation to the group means, using several points rather than a single point, a mean value for the points common to two curves is obtained for each curve in terms of deviation values from the mean of the group in a unit normal curve. Each of these two mean values is multiplied by the standard deviation for the group in terms of the basic units. The difference between these two products is an estimate of the distance between the means in terms of the basic units. Here also the most accurate estimate would be obtained by computing an average based on each of the raw-score intervals covered by both curves. As in estimating the ratios of the standard deviations, it appears likely that five or six points will provide a good first approximation.

Many scaling procedures stop at this point. The means of the curves have been located and their standard-deviation units can all be placed in terms of such basic units as are selected. For any given point on the new scale a precise statement can be formulated as to what proportion of each distribution can be expected to fall below it. Curves can be plotted and conversion tables from raw scores to basic units prepared.

The goodness of fit of these curves can be tested by using a chi-square test to check on the tendency for the expected number of cases to fall below a specified set of raw-score points not used in obtaining the constants for the fitted curves. This seems especially appropriate where only a few points have been used in the fitting in a situation in which skewed curves have been used. It is not clear what conclusion one draws when the fit is found to be poor—presumably, that the procedure was not adequate or the

assumptions not appropriate. No direct information leading to immediate improvement seems to result.

A more practically useful approach to the problem of testing the adequacy of scaling appears to be to use the results of the steps described above to establish an actual set of basic units corresponding to the raw scores. Using these first estimates of the scaled units, means and standard deviations for each of the groups are calculated. If these means and standard deviations in terms of the tentative basic units are not the same as those on which the scaling to these basic units was based, it is concluded that the estimates were not satisfactory for scaling purposes. New scaled values are obtained corresponding to the various raw scores using the obtained means and standard deviations to calculate the proportions of the normal (or skewed curve) for each group which fall below each of the scaling points in the basic distribution. Means and standard deviations are calculated using these new basic units, and these values are compared with those determining the scale values. This process is repeated until stable values are obtained so that the means and standard deviations obtained in the checking step are the same as those used in determining the proportions of frequencies for the various groups. The new units may then be tested on the distributions of entirely new groups.

In obtaining the equivalent basic units for each raw score, two methods may be used. In the first method estimates of the equivalent values are obtained from each of the groups. Because of sampling fluctuations, lack of precise agreement of the data with the basic assumptions, and similar disturbing factors, these estimates will ordinarily not be identical. It is, therefore, necessary to combine them in some way to arrive at a single value. It would seem appropriate to weight the values inversely as their variance errors in terms of the basic units. However, this neglects the varying accuracy with which their means and standard deviations have been established. This again is not so important if an iteration procedure is used. The second method is to cumulate the expected frequencies below each of the basic unit values being scaled for all groups. This over-all frequency is then used to obtain the corresponding raw score from the original combined distributions. This provides a rough weighting in terms of frequencies. It tends to obscure discrepancies in scale values for the various groups, although these will usually appear in the comparison of the new means and standard deviations.

As indicated at the outset of this discussion, the crucial test for any scaling procedure designed for practical use is: will it distribute new groups of the specified type in accordance with the form of curve used to establish

the units? A satisfactory finding from a series of chi-square tests of goodness of fit for new groups would seem the best possible evidence in favor of a particular scaling procedure. A final word of caution should be included concerning fitting curves with small numbers of points. Especially if skewness is allowed to vary, radically different results may be expected from very small sampling fluctuations. For example, in a recent study the final fitted values as determined from four selected points for a pair of curves were found to give skewnesses of $+0.18$ and -0.04 respectively for the two curves with a ratio of 1.02 for their standard deviations. However, values for the skewnesses of -0.50 and -0.85 respectively and a ratio of the standard deviations of 0.87 were also such a close fit that a change of 0.3 of 1 percent in the frequency for one of the distributions below one of the four points used would have provided a perfect fit with these quite divergent values. This does not suggest great stability for the obtained values in a practical situation. Although four points are sufficient to determine the curves, sampling errors are such that the constants for the pairs of curves are going to be quite different if different points are used.

Possible limitations of the method of normalizing overlapping distributions

Theoretically, it is always possible, for *any* single continuous distribution, regardless of its form, to convert the original scores into a new set of scores that will be normally distributed. Therefore, using this method, no checks and no warnings are provided the user. The simultaneous normalization of two or more distributions, however, assumes that certain relationships exist among the original distributions. If these relationships do not hold, the user finds out in trying to apply them and is aware that the procedure is inapplicable.

It is self-evident that it is impossible to construct a single conversion table for every possible pair of raw-score distributions for the same test that will simultaneously normalize both of the distributions of converted scores. Suppose, for example, that two distributions of raw scores on the same test are as pictured either in the left or right half of Figure 66. Quite obviously, in either the left- or right-hand figure, any conversion table that would make the *A*-distribution more nearly normal would inevitably make the *B*-distribution depart even farther from normal.

It is, of course, possible that many pairs of two distributions, both markedly non-normal, but with different means and standard deviations, can both be made normal by using a single conversion table. However, it is not known, for the typical achievement test, *to what extent* the varia-

tions in form of distribution actually found from school to school are of this type, or to what extent they are of the type illustrated in Figure 66, which cannot be "ironed out" by *any* single conversion table.

Selection of the populations for which norms are to be established

A test "norm" is an estimate of some characteristic of the distribution of scores for a specific population. A major purpose of test norms is to facilitate the interpretation of an individual's scores on the test by making possible a comparison with the scores made by others. The meaningfulness and dependability of a norm with reference to this purpose depend primarily

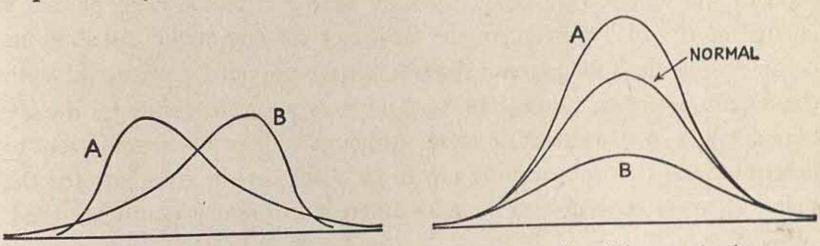


FIG. 66.—Possible forms of raw-score distributions for different schools

upon (1) the exact nature of the population for which the norm is established, (2) the number of pupils and schools selected from this population to establish the norm, (3) the degree to which the sample selected is representative of the specified population, (4) the nature and uniformity of the conditions under which the test was administered to this sample, and (5) the extent to which the score being interpreted was obtained under these same conditions.

Unfortunately, for some of the published tests for which norms have been provided, no clear description has been provided of the population for which they were established. Considering the none-too-well-known fact that school differences are sometimes so large that the lowest-scoring students in one school may score above the highest-scoring students in another school on the same test, a failure to specify for what kind of schools a test norm was established renders it almost completely meaningless.

The definition of the population for which a norm is to be established is, of course, highly arbitrary. One might, for example, try to establish a norm for an elementary school achievement test for all seventh-grade pupils in the United States. However, the variations in achievement from one part of the country to another, or from one type of school organization to another (for example, from ungraded one-room rural to city junior high schools), or from one type of community to another, are so great as to make such highly generalized norms very difficult to establish. In gen-

eral, the more homogeneous the population for which the norm is established, the more meaningful the norm, granting that the individual whose score is being interpreted naturally belongs to this population. For purposes of educational guidance, particularly, it is much more meaningful to know how a student's performance compares with that of other students who have had very nearly the same learning opportunities, than to know how it compares with the performances of a heterogeneous group of students who have had a wide variety of learning experiences and opportunities.

What is needed, ideally, is a series of norms established for a large number of homogeneous groups, so that each test user may make comparisons with a variety of norms appropriate for his purpose. To mention only a few possibilities for, say, an elementary school arithmetic test, one might have one table of norms for seventh-graders in large industrial communities, another for seventh-graders in ungraded one-room rural schools for Negroes in the Southern states, and another for schools in small agricultural or residential communities in the Middle West where the traditional 8-4 organization is still in vogue. Again, there might be one set of norms for New England, another for the Southern states, another for the Pacific states, etc. Or there might be separate "ability" norms for pupils at each level of intelligence within each grade. There is, then, a very large number of populations for which a norm may be established, and in which a large number of school people would be interested.

It is clearly impracticable to provide with any test enough norms of this highly specific character to satisfy all possible demands. The tendency in standardized test construction, therefore, has been to go to the other extreme and to provide only a single highly generalized norm. The development of such norms presents a problem since it is difficult to provide an exact and generally acceptable definition of the population for which a nation-wide norm should be established. Few standardized tests are provided with meaningful descriptions of the exact nature of the sample upon which their norms were established, or of the conditions under which the pupils in those samples were tested. Depending upon how the sample was selected, however, the so-called nation-wide norm on a test may be either "high" or "low." It is well known that norms on current tests differ markedly in this respect. It has been found, for instance, that when several independently constructed standardized tests in arithmetic are administered to the same group of pupils, the average score of these pupils may be "at" the norm for their grade on one test, one or two grade levels above the norm on another test, and one or two grade levels below the norm on still another test. "Nation-wide" norms, then, even if established on a de-

pendable and describable basis, would fit very few of the diverse populations in which the tests might be administered.

In summary, an extremely important feature of a test norm is that the population for which it is established be very clearly described to the test user. A test intended for widespread use should be provided with multiple sets of norms, one for each of a number of relatively homogeneous populations. Finally, it is extremely important that the norms be established on a sample that is truly representative of the population described. The difficulties in meeting this latter requirement are discussed in the following section.

Sampling problems in establishing norms

Individuals who have received a limited amount of training in sampling theory are apt to accept large numbers as a guarantee of adequate sampling. Anyone familiar with the formula for the standard error of the mean knows that it requires only a random sample of a few hundred students to justify a high degree of confidence that the obtained mean is not in error by more than a tenth of a standard deviation. This knowledge seems to bring with it a tendency to feel that if the sample includes several thousand, rather than several hundred, students, it must certainly represent the total group very accurately. This is, of course, frequently not so. This situation is not the result of defects in sampling theory or any special conditions regarding either schools or educational achievement which invalidate the sampling procedure. It is due solely to the fact that the several thousand students on whom the norms are based are ordinarily not drawn from the total population in a way even remotely resembling an appropriate procedure for obtaining a random or representative sample of all students.

This is due primarily to the fact that students are found in classes, schools, and school systems which differ very markedly from one another in the achievement of their students. A sample of even 100,000 students from one school system is likely to be unrepresentative of other groups, since it is selected on a basis which is not independent of school achievement. For the purpose of getting a really random sample of a few hundred students from the entire population of schools, we cannot regard this one school system as contributing more than a very few cases. A sample of 100,000 from this system may be almost precisely representative of the achievement scores on the test in this one city. However, it would provide very little information about what scores to expect from other cities.

Stated in other terms, a large part of the variability of scores on educational tests is associated with the particular school in which the individual is found. Variability certainly exists among the students in a given

school. However, the sampling procedure must give all schools an opportunity to be represented in the sample. In general, it may be said that in current practice large numbers are not a problem in obtaining norms for educational tests. What is needed is norms based on really random or representative samples from the specific populations they are intended to represent.

If the population of schools involved may be divided into homogeneous subgroups, such as urban and rural schools, or schools in different enrollment classes, or schools in different geographic areas, then the number of schools in the standardization sample should be proportionally representative of these subgroups. In cases in which it is not possible to obtain adequate representation from all known stratifications within the population, it is desirable to weight the values for the schools used in accordance with the size of the group which they can reasonably be considered to represent. This procedure is illustrated in the Norms Booklet of the Cooperative Achievement Tests published in 1938. This practice is not used as often as it should be by test constructors.

The need for norms of school achievement

One of the major uses claimed for educational achievement tests is in *evaluation* of the content and organization of the curriculum and of the quality or effectiveness of instruction in individual schools. This implies an evaluation of group, rather than of individual pupil, performance. What is needed for this purpose, obviously, is a type of norm that is descriptive of distributions of measures of group achievement, that is, of school averages, rather than of distributions of individual pupil scores.

Since such norms are rarely provided, what the school principal does is to attempt to interpret his grade averages or grade medians in terms of the norms that *are* provided—norms descriptive only of distributions of pupil scores. What he really does, often without knowing that he is doing so, is to assume that the school averages are distributed just as are individual pupil scores. School averages, of course, are much less variable than pupil scores (although not as much less variable as many may assume), and he therefore makes wrong interpretations of his results. A school average that coincides with, say, the 75th percentile in the pupil score distribution may be a remarkably high average, but since it is "at the 75th percentile," it might not be regarded as at all high by the principal. What he does would not be so serious, however, if there were a constant relationship between the variability of school averages and the variability of pupil scores. In that case he could at least tell on what tests his school compared most

favorably with other schools, even though he would be misled as to the true relative performance in any test. Actually, however, this relationship is not constant, but varies considerably from subject to subject and from one grade level to another.

The situation may be summarized as follows. School averages are less variable than pupil scores on all educational achievement tests (but much more variable than the means of truly random samples of the same size drawn from the whole population). The relative variabilities of school averages differ markedly from test to test and grade to grade. If school administrators are to interpret properly measures of school achievement for purposes of evaluation in general, they must be supplied with norms specifically adapted to that purpose; that is, they must have norms descriptive of distributions of school averages for defined populations of schools.

Fundamental point of reference

It is very desirable to facilitate the interpretation of test scores by giving them as much direct meaning as possible. It is always necessary to supplement the knowledge inherent in the scores with other normative data. However, if much information is built into the score itself, continual use makes its interpretation more and more direct and immediate. It is also of great assistance if such fundamental built-in meanings can be as constant from one test to the next as possible.

The most basic consideration in establishing a system of scores for rendering scores comparable for different tests is the selection of a fundamental point of reference. It might seem that the most desirable procedure would be to establish a zero point for each test. However, this is extremely difficult for most educational tests, and it would add but little to the use of tests for most purposes.

There are many other points of reference which may be selected. In defining his original T-scores, McCall (*15*) used the average performance of unselected twelve-year-olds as the basic point of reference. Although this reference point has certain advantages, it does not have satisfactory stability because of differences in age at school entrance in different regions, and also because of differences in policies with respect to passing and failing. Since students are classified by grades and years of study, reference to the performance of such groups is more meaningful to school officials than reference to an age group. Another point of reference which has been used is the average performance of adults. This point, however, is greatly influenced by such a factor as the progressively greater number of years during which the average individual attends school.

A point of reference which is especially well adapted to tests for use at the secondary and college level is the score which the average white child in the United States would make at the end of the particular course if he attended an average school and had had instruction in the subject in question which was typical with respect to quality, amount, and grade placement.

This concept has the advantage of referring all tests to the same basic population. The point of reference would be similar for every subject—French, English, or anything else—regardless of the fact that in many schools all students who take French are on the average much superior to all students who take English (a much larger group). This definition of a fundamental reference point is especially desirable because of the wide differences in placement and selectiveness among various courses in local school systems throughout the country.

The advantages of such a point of reference are obvious. It incorporates the "national norm" directly into the individual scores and makes their interpretation with respect to this point simple and direct. It should be noted further that this definition stabilizes the norms. Norms which are defined in terms of the quality of the groups who happen to be taking the various subjects in a particular year are bound to fluctuate from year to year, no matter how adequate the sample. For instance, some high school norms established only a decade or two ago are not satisfactory at present since the number and character of the students enrolled in secondary school has changed markedly. Also, various factors, such as the temporary dominance of some educational doctrine or the alteration of college entrance requirements, may change very greatly the type of individual enrolled in particular subjects. From the point of view of individual guidance, the lack of comparability in the present norms from subject to subject is a greater deficiency than the failure of these norms to be representative of current groups. It is very difficult to get a clear picture of an individual's relative achievement in various subjects if the scores are not directly comparable with respect to some well-defined basic group. Even though the information necessary for a precise comparison between subjects is available, the teacher can scarcely be expected to derive more than a rough idea of the relative performance of a pupil on tests in various subjects, because of the complexity of the process of adjusting the scores.

Assuming that a system of scores scaled with reference to points such as those just discussed is desirable, the possibility of obtaining such a system may still be questioned. Obviously, not all the white children in the United States are enrolled in even one of the subjects involved. Furthermore, most of the students taking certain secondary school subjects, such as trigonometry, are very highly selected. However, many students who appear

to be representative of this "average child" do continue in school, and some of these will be found enrolled in each of the various subjects. The method does not break down even if it is impossible to locate an adequate number of average individuals, since for such a selective subject the bridging or equating may easily be done at some higher defined point established on a similar test which then serves as an anchor test. For example, the basic point of reference for Latin might be determined by equating with French tests. (See the discussion on equating in a later section of this chapter.)

To establish beyond a doubt that any "standard group" of secondary school students is truly representative of white children in the United States is obviously impossible. Geographical location, size of community, economic status, specific abilities, interests, attitudes, and many other factors would have to be controlled. Fortunately, however, achievement in the secondary school is dependent primarily on only a few of these factors and the "standard group" need therefore be representative only with respect to these few important ones.

The general aptitude of the students, their attitudes toward schoolwork, and the characteristics of the teaching within the school system are generally accepted as the most important determiners of achievement in the secondary school. The most useful criteria for defining the average or typical white child in the United States for the purpose of establishing a fundamental point of reference are probably the following: (1) score on a well-standardized comprehensive achievement test administered in one of the grades before much selection has occurred; (2) score on a well-standardized test of general academic aptitude or intelligence; (3) grade placement for age; the modal age-grade group has many advantages for such purposes.

In establishing the representativeness of high school groups, the results from an elementary school achievement test, when compared with aptitude or intelligence test results which emphasize general intellectual development rather than school development, should be valuable in indicating the attitudes of the group of students toward schoolwork and the characteristics of the teaching within the school system. These results would also provide information concerning the representativeness of the group with respect to the intellectual capacity of the students. The intelligence tests should maximize the opportunity to show the effects of general out-of-school learning.

In practice it is found more accurate to use "equating" procedures than simply to obtain the mean score of the comparatively small group of individuals who fulfill all the requirements set, although these procedures do not avoid certain basic difficulties. These "equating" procedures will be discussed in the section on "Comparability."

Size of units

In order that scores have as much similarity as possible to simplify interpretation, the size of the units in which they are measured must be similarly defined for tests in all subjects. It is further desirable that this size have a meaning which would be useful in interpreting the scores and that the fineness of the units be of the order of magnitude most appropriate to express their accuracy of measurement.

With respect to the question of meaning, it would be possible to give the unit a size so that some multiple of it such as 10, 20, or 50 represented the average gain produced by a year's study. How unsatisfactory this procedure is at the high school or college level has already been pointed out (page 707). Another procedure which appears to be more desirable is that of making the unit depend on the variability of the scores in a defined group by giving it some value such as a tenth of the standard deviation of the distribution of test performance in that group. This alternative has two advantages. It provides for units which have similar meaning in various subjects in the sense that they are based on the variability of the defined or "standard" group in performance on the several tests. They are, therefore, readily usable for detecting the strong and weak points of an individual or a group in relation to the scores of the "standard" group. Secondly, units based on group variability are more stable than those based on the means of groups at successive levels, especially at the higher levels. The standard deviation has a smaller standard error than that of the difference between two means.

With respect to the second question, which is that of the degree of fineness desirable in the units, the considerations are ease of handling, correctness of interpretation, and the retention of as much precision in reporting scores as is useful. The standard error of an individual's score on a test for which the reliability coefficient is 0.96 would be 0.2 of a standard deviation. Very little information would, therefore, be lost in most cases if the scores were reported in units of 0.1 of a standard deviation. Furthermore, such units would make it possible to express scores on practically all tests in two place numbers since this would allow a range of ten standard deviations. It should be clear that reporting a score as being 627 when the measurement is so inaccurate that the best statement that can be made is that the chances are about two to one that the individual's "true" score would be found to be between 587 and 667, is of exceedingly questionable utility if not positively misleading to those interpreting the score.

For some purposes a system of single-digit scores, such as the stanine scores described on page 727, may provide the most useful score units.

Single-digit scores provide a maximum in the way of simplicity and ease of handling and interpretation, but in general are too coarse to preserve all of the information contained in the raw scores on educational achievement tests. In cases where it is believed that the accuracy of measurement justifies greater refinement of units for a specific purpose, this may be achieved by dividing each unit into three equal parts with the aid of the symbols "+," ".", "—". These symbols can easily be included in the remaining three spaces on tabulating cards having twelve rows to the column. In such cases they are much more readily handled than two-digit numbers. Although they do not lend themselves well to adding and calculating machines, these symbols can ordinarily be omitted without serious loss in calculating statistical values for groups.

Comparability

A fundamental necessity in any test of which more than one form is published is *comparability* between the scores obtained from the several forms. There are a few exceptions to this rule, such as tests whose sole purpose is to place the individuals taking a single form in rank order from highest to lowest, and diagnostic tests whose purpose is to determine only whether the specific items have been mastered.

There has been much confusion as to the meaning of the term "comparability," and it appears desirable that more precision in terminology be obtained. As the term is used in this discussion, scores on two tests are *comparable* for a given population if the two distributions of the "true" scores for these tests are identical for any (and all) large groups selected from this population. If the reliability of measurement is the same for the two tests for the population involved, then similar results will be obtained if the distributions of obtained values are compared. Since "true" scores are never available, a more practical definition may be that scores on two tests are comparable for a given population if their two mean scores are very nearly identical for any (and all) large groups selected from the population. The usual practical procedure for establishing comparable scores for tests having similar reliability coefficients is to make the distributions of obtained scores identical for each of several large groups selected from the population involved so as to cover the complete range of scores. If the reliability coefficients differ substantially, *estimated distributions of true scores* for these groups must be substituted. (These are *not* the same as the distributions of *estimated true scores.*) This definition, it should be noted, applies to scores from different tests as well as to scores from different forms of the same test.

As mentioned in the introductory section of this chapter, comparability which would hold for all types of groups—that is, *general* comparability between different tests, or even between the various forms of a particular test—is strictly and logically impossible. Unless two items are identical, the proportion of individuals responding to them correctly cannot be expected to be the same for varying groups. For example, two forms of a test might be constructed so that they gave practically identical distributions of scores for 50,000 students tested in New York City. However, when these same two forms were given to another group of 50,000 students in Los Angeles, the distributions for the two forms might be found to be quite different. One of the supposedly comparable forms might conceivably have a greater proportion of items which were emphasized in the Los Angeles course of study, and which would, therefore, be quite easy for that particular group, than did the other form, even though parallel items in the two forms were of equal difficulty for the New York group because they were given equal emphasis in the New York City course of study.

It should be emphasized, however, that, as a practical matter, comparability of scores from forms of the same test can be approximated quite closely if proper methods are used in the construction and equating of the examinations. Such methods of construction include a detailed balancing of the content of the various forms with respect to various types of item classification, such as topics included, operations required in responding, and the amount of time necessary for reading and answering the item.

In this discussion, "comparability" has been used in a much broader, though in certain respects more definite, sense than has usually been implied when the term "comparable" forms has been used. The various forms of a given test which are represented by the publishers as essentially interchangeable have been variously designated as "equivalent" forms, "similar" forms, "parallel" forms, or "comparable" forms. The scores from these forms are, however, not always *comparable* as defined above. On the other hand, it is possible for tests in somewhat different subjects to fulfill the requirements mentioned and thus have comparable scores. Therefore, in describing the forms of a single test which are presented by the publishers as interchangeable in function, this discussion will use the term "similar forms." Forms which are truly interchangeable in that they measure the same functions with equal accuracy and are reported in comparable scores will be designated as "equivalent forms."

Because of the usage of terms mentioned above, there has been some confusion between the concepts of similarity in the distributions of scores for large groups on two forms of a test and similarity of an individual's scores on these forms. Since tests are essentially sampling devices, an in-

dividual cannot be expected to obtain identical scores on similar or even "equivalent" forms. One sample will usually correspond somewhat more closely to the specific character of his training, experience, and aptitudes than does the other. The extent to which each of the several similar forms will place the individuals in a given group in approximately the same relative position along the scale represented by the test is a measure of the consistency or reliability of the test. If the degree of consistency of the forms of a particular test is low, the scores on the similar forms may be very different for an individual even though they are expressed in comparable scores. The degree of consistency of the forms of a test is usually reported in terms of the reliability coefficient of the test or the standard error of measurement of a test score. These terms are discussed in chapter 15.

Constructing equivalent forms

One of the most frequently used techniques for securing similar forms of a test has been to construct the forms originally so that their *raw* scores tend to be directly comparable. To construct such forms, a large number of items is tried out experimentally, the difficulty (and sometimes the discriminating power) is determined for each item, and then items are selected for the various forms such that they simultaneously show the same distribution in the various categories of the table of specifications, the same distribution of item difficulties, and (sometimes) the same distribution of discriminating power for the various forms. This procedure, carefully followed, can produce forms whose raw scores are at least approximately comparable. This method of securing similar forms has usually been less expensive than the methods to be explained in the following section, and the convenience and simplicity of directly comparable raw scores was formerly deemed essential to the wide use of a test. The chief defects of the method are:

1. It is likely to be rather inaccurate, since the "difficulties" of the same item are in many cases not entirely comparable for the tryout form and the finished test. There are two principal reasons for this: first, the position of the item in a test may affect its "difficulty," even though all students attempt all items, because of the tendency to hurry on the later items under usual testing conditions; and, second, items moved from their context in the experimental form to new context in the final form sometimes show changed "difficulty" values.

2. It is frequently impossible to divide the materials from the tryout forms into closely equivalent forms without making too great a sacrifice with respect to the sampling of content in each of the forms. That is, it is

difficult to secure identical distributions of item difficulty and at the same time to maintain the same distribution of emphasis on the various topics or abilities tested. This is especially true in tests involving indivisible blocks of items, such as reading paragraphs accompanied by several questions.

3. It is difficult to make improvements in later forms of the test, since the number of items in each part of the test, the general character, and the length of the items must be kept the same as in the previously published forms.

4. The raw scores are comparable under only one particular set of conditions of administration. For example, raw scores obtained on tests administered with separate answer sheets are not generally comparable with those obtained from the same tests when separate answer sheets are not used, but tend to be somewhat lower than raw scores obtained on tests in which the choices are indicated in the test booklet.

5. Even though the distributions of item difficulties in the two forms were identical, the items in one form might be more closely related to each other, that is, more "valid" in the internal consistency sense, than those on the other form, causing a greater spread in the scores on the first form since a person getting one item right would tend to get many others right and vice versa.

Ways of securing comparable scores for different forms or tests

To obtain comparable scores for forms of tests which have not been constructed to yield equivalent raw scores, the scores are "equated." A common method of obtaining the data necessary for equating scores is to administer the two or more forms to all of the individuals in a particular group. To equalize practice effect and control administrative conditions, half of the group takes the first form first and the other half takes the second form first. If the test is very long, or if more than two forms are involved, this technique often becomes impractical. In this case the procedure usually used is to administer each of the several forms to matched individuals within various groups or to random halves of various groups. With this procedure, it has usually been found best to equalize such administrative factors as timing, testing conditions, and instructions by having the two or more forms administered simultaneously to the two halves of the same group. This is very simple when, as is usual, the forms have the same directions for administering and the same time limits. Various statistical methods have then been employed to obtain comparable or equivalent scores for the forms being compared. These are discussed in the following sections.

1. *Comparability by equating means.* One of the earliest procedures for

obtaining comparability was to compute the means for the two distributions of raw scores, and if they did not differ more than could reasonably be attributed to sampling fluctuations, they were called "comparable" or "equivalent." If the difference was significant, it was suggested that an amount equal to this difference in means be added to or subtracted from a score obtained on one form to make it "equivalent" to that obtained on the other form. This procedure provided only a rough approximation, since it was not always true that the difference between forms was uniform throughout the range of scores. The two forms frequently differed also in their variabilities and in the shapes of their distributions.

2. *Comparability by use of regression techniques.* A procedure which was regarded as an improvement on the matching of means was to compute the correlation between the two forms and obtain by means of the usual regression formula the "best estimate" of the score on one form, with the score on the other known. A variant of this procedure was to compute the best estimate of a third and possibly more valid measure from each of the two forms. It should be clearly noted here that such a procedure does not give *comparable* scores. The "best estimate" of a person's score on form A, when it is known that his score on form B is 66, may be 73. But it does not follow that 66 is the best estimate of a form B score, when it is known that the score on form A is 73, unless these scores are at the mean of the group or the correlation between the two forms is very close to 1.00. Best estimates depend not only on the relative difficulty of the two sets of materials, but also on the degree of relation between them. Thus, if the degree of relationship is low, the best estimate of the score on form A which can be obtained from the score made on form B is very close to the mean score on form A. In this situation, the discrepancy between form A scores estimated from form B scores and the corresponding form B scores estimated from form A scores is very large for scores not close to the mean. It immediately becomes evident that best estimates have no universality, but are markedly affected by factors in the particular situation. A table of "equivalent" scores may be prepared using the best estimates from each of the forms of some more valid measure of the particular trait. However, even these corresponding scores are usually not truly "comparable," as will be shown in the following paragraphs.

One weakness of this "regression line" procedure is that if the correlation coefficients with the third, more valid, test differ for the two forms, the results will depend on the level of the equating group used. A group in which the average score is 50 will give a different set of "equivalent" values than will a group in which the average score on the same form is some other value, such as 60. This is due to the effect of "regressing" the

scores toward the mean in obtaining best estimates. An extreme hypothetical case will illustrate this point. Suppose that the correlation coefficient found between form A and the more valid measure, form O, is 1.00, and the corresponding coefficient for form B is 0.50. Then if the average scores on all three of the tests are 50 and the variabilities are equal, the best estimate of the score on the more valid measure, form O, for a person with a score of 60 on form A is 60. Similarly the best estimate of the form O score for a person with a score of 70 on form B is 60. Thus, 60 on form A is "equivalent" to 70 for form B. If, now, a group is used for "comparing" these tests for which the average scores on the three tests are all 60 and for which the variabilities are equal, the best estimate of the score on the more valid measure, form O, for a person with a score of 60 on either of the tests is 60. The conclusion in this case would be that 60 on form A is "equivalent" to 60 on form B. These contradictory conclusions reveal the weakness of such regression line methods in this type of situation.

A more important weakness of this regression procedure, which it has in common with the typical standard score method, is described below.

3. *The standard score method.* The standard score method, also called the "line of relation method," consists essentially in making a uniform adjustment in the scores in accordance with the difference in the mean scores for the two forms, as in the first method mentioned above, and, in addition, adjusting the scores by multiplying by a constant number so that the variability, in terms of the standard deviations of the two series of scores, will be the same for the equating group. In this procedure the mean is frequently given some standard value such as 0, 50, or 100 and the standard deviation is made equal to 10 or 20 or 100. The weakness of this procedure is the assumption of similar shapes for the distributions and a straight line relationship between the two series of scores. Often one form as compared with the second has a few more very easy items, a few less moderately easy ones, and then a similar distribution for the balance of the scale; or some other situation may exist which produces a definitely curvilinear relationship between the two series of scores. If the reliability coefficients for the two tests differ as described in the previous paragraph, the pairs of scores considered to be equivalent are more appropriately those which yield equal scores in terms of deviations from the means when multiplied by the square roots of their respective reliability coefficients. (See following paragraph.)

4. *The equi-percentile method.* For the reason just noted, and for various others which have been previously mentioned, it appears that the most satisfactory method of obtaining "comparable" scores for the various forms of a given test or for different tests, is based on the proposition that "true"

scores which are of "equal difficulty," that is, "true" scores which would be exceeded by equal proportions of the group, are "comparable." In using this procedure, the proportion of the "true" scores of the group which would fall below a certain value on one form is calculated and the score on the other form below which an equal proportion of the "true" scores on that form would fall is listed as comparable to this particular value. It is obvious that "true" scores are never available for individuals. "Estimated true" scores can be calculated for each person, but it is emphasized that in this discussion reference is *not* made to distributions of "estimated true" scores. Rather reference is made to estimates, based on the obtained score distribution, of the distribution characteristics (such as M and σ) of the hypothetical "true" scores which would be obtained by averaging an indefinitely large number of scores from similar forms for each individual. For example, it is well known that the variance of the "estimated true" scores is equal to the square of the reliability coefficient times the variance of the obtained scores. On the other hand, the variance of the "true" scores were they available would be found to be equal to the product of the first power of the reliability coefficient and the variance of the obtained scores. These relationships are symbolically expressed below.

$$\sigma^2_{\text{Est'd true scores}} = r^2_{12}\,\sigma^2_{\text{obt'd scores}}$$

$$\text{Est'd } \sigma^2_{\text{true scores}} = r_{12}\,\sigma^2_{\text{obt'd scores}}$$

Where one is equating scores on forms or tests having similar reliability coefficients, he can work with obtained scores, rather than true score distributions, since approximately the same relation between true and obtained scores exists for both forms.

Instead of preparing a table of "equivalent" scores directly by listing in adjacent columns scores below which equal proportions of the distribution fall, it is customary to plot points on rectangular coordinates corresponding to these pairs of scores. A smooth curve is then drawn through the points and the pairs of equivalent scores are read from this curve. This is called the smoothed equi-percentile method of equating scores. If more than two forms or tests are to be equated, one of these is used as the basic test to which others are equated.

An example illustrating this equating procedure follows. Table 13 presents the distributions of raw scores on forms A and B of a science test for matched samples of 663 and 682 cases respectively. The table gives, for each of a number of score limits for the distribution on form A, the corresponding raw scores for form B. These pairs of raw scores are plotted in Figure 67, the plotted points being indicated by the circles. The last point plotted in the upper right-hand corner represents the maximum pos-

sible scores on the two forms. A smooth curve has been drawn through the plotted points. Comparable scores are then read from this smoothed curve. For example, a score of 69 on form A is comparable to one of 70 on form B.

Experience in preparing a number of sets of "equivalent" scores in this way soon reveals, when subsequent checks are available, that smoothing

TABLE 13

DISTRIBUTION OF SCORES AND EQUI-PERCENTILE POINTS FOR TWO FORMS OF A SCIENCE TEST FOR MATCHED SAMPLES OF TENTH-GRADE PUPILS

Scores	Frequency Form A	Cumulative Frequency Form A	Corrected Cumulative Frequency Form A for N=682	Frequency Form B	Cumulative Frequency Form B	Points on the Score Distribution below Which Equal Proportions of the Scores Fall	
						Form A	Form B
88–90.....	....	....	....	....	....	....	....
76–78.....	....	....	....	1	682	90.0	90.0
73–75.....	....	....	....	1	681	....	....
70–72.....	2	663	682.0	4	680	72.5	78.5
67–69.....	1	661	679.9	2	676	69.5	72.4
64–66.....	4	660	678.9	4	674	66.5	71.7
61–63.....	5	656	674.8	7	670	63.5	67.7
58–60.....	15	651	669.7	12	663	60.5	63.4
55–57.....	9	636	654.2	12	651	57.5	58.3
52–54.....	12	627	645.0	18	639	54.5	56.0
49–51.....	18	615	632.6	26	621	51.5	53.4
46–48.....	16	597	614.1	38	595	48.5	50.7
43–45.....	40	581	597.6	40	557	45.5	48.8
40–42.....	41	541	556.5	50	517	42.5	45.5
37–39.....	48	500	514.3	54	467	39.5	42.3
34–36.....	72	452	465.0	73	413	36.5	39.4
31–33.....	85	380	390.9	82	340	33.5	35.6
28–30.....	98	295	303.4	91	258	30.5	32.2
25–27.....	97	197	202.6	72	167	27.5	28.7
22–24.....	68	100	102.9	55	95	24.5	24.8
19–21.....	22	32	32.9	24	40	21.5	20.6
16–18.....	6	10	10.3	8	16	18.5	16.4
13–15.....	3	4	4.1	8	8	15.5	14.0
10–12.....	0	1	1.0	....	....	12.5	12.9
7– 9.....	0	1	1.0	....	....	9.5	12.9
4– 6.....	0	1	1.0	....	....	6.5	12.9
1– 3.....	1	1	1.0	....	....	3.5	12.9

can result in worse, as well as better, tables of "equivalents" than would be obtained from the distributions without any smoothing. The novice is likely to pay too much attention to minor fluctuations in the positions of the points at the center of the curve rather than maintain a relatively straight line or smooth curve. Lack of experience is also likely to lead to paying too much attention to scattered points at the ends of the distribution which are based on only a few cases. It is usually desirable to connect the maximum

scores possible on the two tests to the points based on a fairly substantial number of cases by means of a relatively straight line or smooth curve in accordance with the trend in the more adequately determined positions. It is helpful to calculate and plot estimates of the standard errors corresponding to a few values next to the appropriate points on the chart.

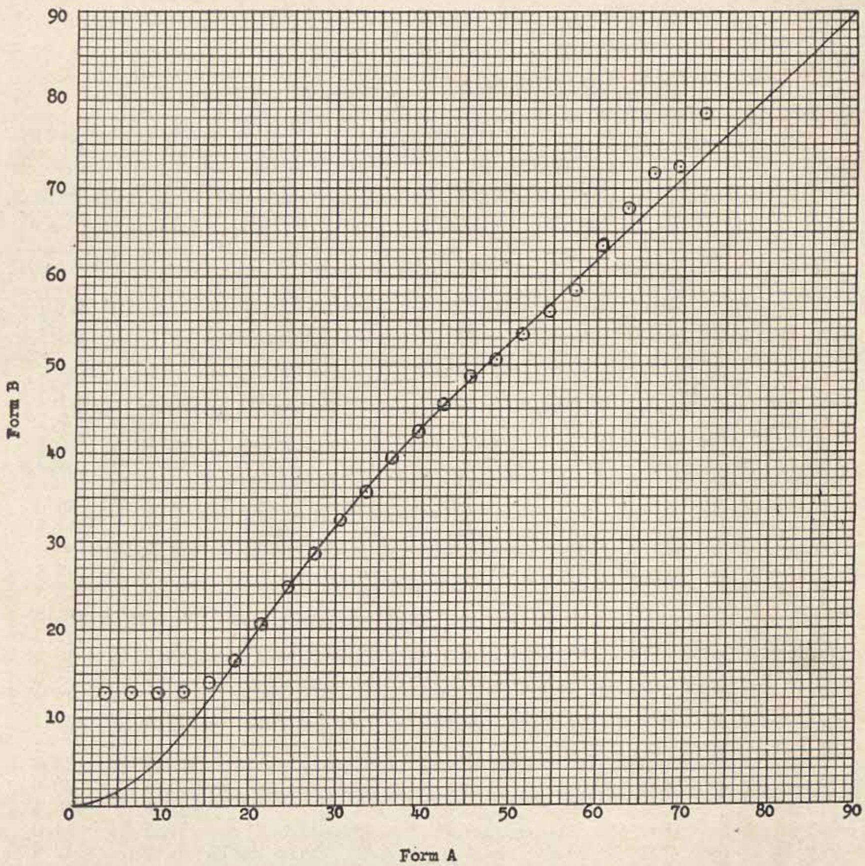


FIG. 67.—Illustration of method of equating scores by equi-percentile curves

Usually some outside information is available, such as that the forms were intended to be of equal difficulty, which can guide the smoothing process. In the case of different tests being equated, this outside information may suggest that the maximum possible scores cannot reasonably be expected to be equivalent in difficulty. The best type of training for this type of smoothing is to divide the equating sample into halves or to obtain two independent samples. After the plotting and smoothing operations have been completely carried out for one sample, the points from the second

sample can be plotted on the same chart. This will tend to provide immediate corrections for systematic errors being made in smoothing. The best estimate for the correct position of the curve will, of course, be somewhere between the two sets of points. Since extrapolation is likely to lead to substantial errors, *it is very desirable to have overlapping groups covering the entire range of possible scores.* For the usual test which covers only a moderate range, 500 to 600 cases are generally regarded as a satisfactory minimum number.

The accuracy of equating is also a function of the representativeness of the equating group's scores on the two forms with relation to the type of groups for which the test is intended. Furthermore, if the same group takes both forms, the order needs to be counterbalanced so that half takes form A first and half form B first to equalize the effects of practice.

It should be emphasized here that this equating procedure, and also the definition of "comparability" which was given earlier in this section, may be readily generalized to refer to "comparable" scores for different tests as well as for various forms of the same test.

In the equating of scores from different tests the need for referring calculations to distributions of "true" scores becomes more evident, since in this situation the reliability coefficients for the two forms are more likely to differ by a substantial amount. An illustration of the way in which "comparable" scores derived from distributions of obtained scores will differ for groups in which the average scores are different, even though all other factors are controlled, will be given. Suppose scores on two tests (A and B) are to be equated, the reliability coefficients for which are 0.64 and 1.00 respectively. The reliability of the second test is taken as 1.00 to simplify the discussion. In practice the results would be quite similar if the reliability coefficient for the second test was as high as .95. If the equating were done with a group which achieved a median score of 40 on test A and 68 on test B, these two scores would be considered equivalent because an equal proportion (half) of the group would exceed them. If data were obtained from a second group, the median scores of which were 50 on test A and 80 on test B, these two scores would also be regarded as equivalent for the same reason. A study of the results from the first group, however, would be likely to indicate that the scores of 16 percent of the group exceeded some such pair of scores as 52.5 on test A and 80 on test B. Using the ordinary procedures, 52.5 and 80 would be called equivalent. This, of course, is contrary to the findings in the second group. If each of the scores in the distribution for test A were replaced by the appropriate "true" score, that is, the average score that the individual would make if given a very large number of similar forms of the test, the standard devia-

tion of the resulting distribution would be about 0.8 as large as that of the original raw-score distribution. In this distribution of "true" scores, a score of 50 would be exceeded by approximately 16 percent of the cases and would be equivalent to a score of 80 on test B, since the "true" score distribution for test B would be the same as the raw-score distribution.

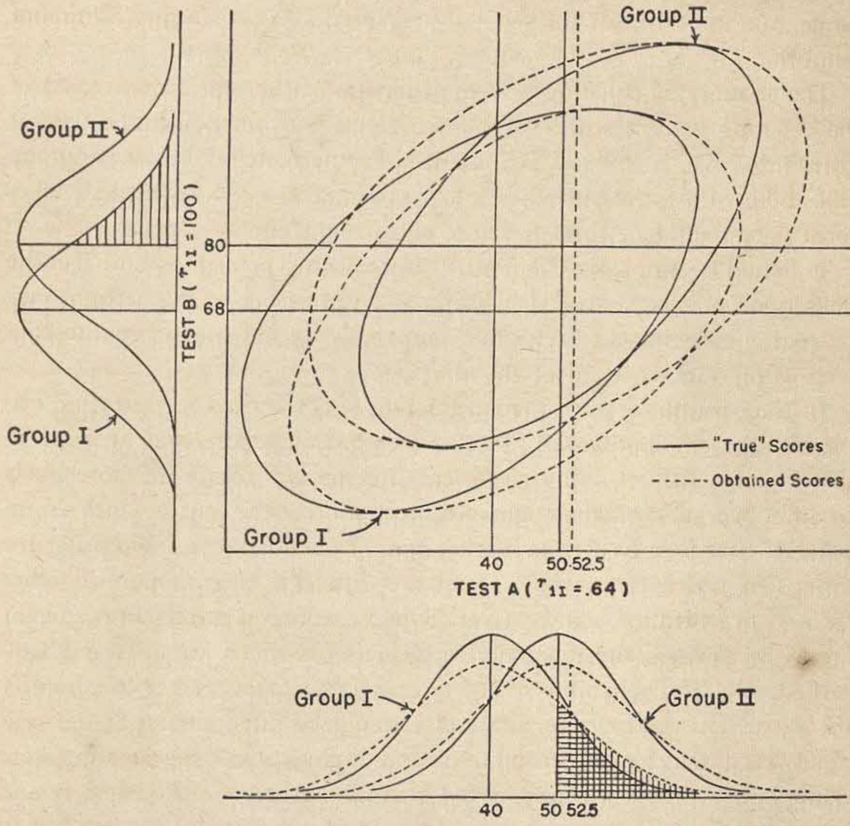


FIG. 68.—An illustration of the procedure for obtaining equivalent scores by the equal proportion method.

The situation outlined above is illustrated in Figure 68 which is explained as follows: The pair of dotted and solid ellipses which extends lower and farther left in the figure represent the scatter-diagram showing the relation between the scores made by group I on test A, as indicated along the horizontal axis, and scores made on test B, which are indicated by position along the vertical axis. The dotted ellipse indicates the relation between the obtained scores from the two tests and the solid ellipse the relation between corresponding "true" scores. The ellipses in the upper right-hand part of the chart indicate similarly the relations between the

scores of the individuals in group II. The frequency curves for the distributions of the scores of the two groups on test A and test B are shown at the bottom and at the left of the figure. It is seen that on test A, for which the reliability coefficient is 0.64, there are two frequency curves for each group, one showing the distribution of the obtained scores by means of a dotted line and another indicating the distribution of "true" scores by means of a solid line. The spread, and, therefore, the overlapping of the two groups in terms of obtained scores, is seen to be greater than in terms of "true" scores. In the case of test B only one curve appears for each of the groups because the distribution of obtained scores coincides with the distribution of "true" scores when the reliability coefficient of the test is 1.00, as in this case. All of these frequency distributions are symmetrical about the lines drawn through the medians as shown.

The 16 percent of the "true" scores in group I which exceeds 80 and 50 are shown by horizontal shading lines on the frequency distribution. The 16 percent above 52.5 on the obtained score frequency distribution of group I on test A is shown by vertical shading lines.

If the reliability coefficients for the two forms being equated are quite similar, the error introduced by using obtained score distributions in place of the theoretically correct "true" score values is very small.

It should be recognized that comparability obtained by any of these methods is only a relative matter. If differences exist between the several forms of a test, they are not measuring exactly the same thing. These differences in the thing measured are usually ascribed to error, or to specific, chance, or sampling factors. The "true" type of behavior being measured is defined as the general factor in a very large number of forms of the test when administered to an exceedingly large and heterogeneous group. The equivalence of the forms of a test is seen to be dependent on the ability of the group tested with respect to the various specific factors in the several forms as well as on their ability with respect to the "true" type of behavior which is measured.

The comparability of the scores for forms of a test must, therefore, always be regarded as relative to the groups used in obtaining the "equivalent" scores. While not invalidating the general procedure for equating forms, this certainly emphasizes the need for a large and representative group on which to base comparable scores. It also suggests that for some tests not one but several tables of equivalent scores would be desirable. A good example of this is the equating of elementary and advanced forms of a French reading test. The comparable scores for one-year students are not usually the same as those for two-year students. The one-year students do relatively better on the elementary form and vice versa.

The foregoing discussion applies also to the problem of obtaining comparability between the scores of tests in different fields. In this latter case it is usually possible to administer the various tests to the same students, thus eliminating to a large extent the problem of matching or sampling. However, in this situation the very large specific factors in the several tests require much wider and more representative sampling than for forms of the same test. The determination of scores of comparable difficulty is consequently much more definitely relative to the particular groups used.

There are a few practical considerations in equating test forms which should be mentioned. If only two short tests are being equated, all students can usually take both forms as noted above. When a series of tests is being equated, it is desirable to give the forms to random samples of students in each of the subgroups used. The accuracy of the equating can be slightly increased if the students in each subgroup are arranged from best to worst, and the forms are passed out in a fixed order. If this is done, it is important to use some device for insuring that the top student of the class does not always receive form A, the next form B, and so forth. This can be easily handled where several groups are being used by making sure that the tests are arranged with a different form on the top of each pile.

These procedures require that larger samples be used in the equating procedure, but avoid the troublesome problems introduced by the effects of practice. Just how much larger these groups need to be will vary with the specific circumstances. The use of random halves of homogeneous overlapping groups has been found in practice to introduce sufficient correlation between the scores of paired individuals so that the same statistical estimates of equating accuracy can be obtained from groups of about twice the size necessary if each individual takes both forms. Thus 1,000 to 1,200 students must take each form to give the same accuracy as 500 to 600 students, all of whom take both forms.

Where a series of tests are to be equated two at a time as in the case of annual forms, a problem arises as to whether successive forms should be equated or all tests should be referred to a common "anchor" form. It can be shown statistically that equivalent scores between the several similar forms will be most accurately determined in the case where all forms are equated to a single anchor form. There are some exceptions to this rule which should be noted. In cases where there is a gradual shift in the content of the test forms, the specificity of the equating to the particular groups used may become a serious problem. If, for example, the anchor form of a language test emphasizes grammar and subsequent forms place gradually increasing emphasis on reading ability, the equating may become quite erratic from year to year, since it will depend to a significant extent on the relative status

of the groups used with respect to reading and grammar. In this situation, the equating of successive forms, between which the changes are relatively small, would provide more generally useful tables of equivalent values.

In equating various other types of tests, the situation is analogous. An anchor form is theoretically preferable under ideal conditions, but in a practical situation more broadly useful results can often be expected by equating tests which are similar to each other rather than to a common anchor form containing content somewhat dissimilar to that of the tests in question.

The specific nature of comparability

In concluding this discussion of comparability, the fact deserves stressing again that "comparability" of two or more tests is always specific to a certain population, and that scores which are equivalent for one population may be far from equivalent for another. For example, one might apply the procedures here described to the pupils in a certain school system in order to determine equivalent scores on an English and a science test. If, then, the same scaling or equating procedures were applied for the same tests to pupils in another school system, quite different results might be obtained, even though the tests were equally reliable for both populations. Still different results might be obtained for a sample representative of both systems, or representative of a larger population to which both systems belonged. Accordingly, the shape of the profile of a set of scores for a given pupil depends upon the population used in equating the tests.

This raises the extremely difficult question of which reference population is most appropriate for a given purpose. Suppose the same battery of achievement tests is administered to all high school pupils in the states of Louisiana, Texas, and Oklahoma. For the interpretation of test results of Houston, Texas, pupils, should the tests be scaled or equated for Houston pupils only, or for Texas pupils only, or for all three states together? For many of the purposes of educational guidance, it would seem desirable to scale the tests for the most homogeneous population to which the pupil in question naturally belongs—in this case the population of Houston high school pupils. For purposes of school evaluation, the tests should perhaps be comparable for the homogeneous population of *schools* to which the given school belongs. In practice, it is not feasible to scale the tests for each of the many homogeneous subpopulations that may employ the tests, and the usual procedure is to scale the tests only once and for the most inclusive population of test users. The resulting problems of test interpretation are very difficult ones and demand a thorough appreciation of the factors involved on the part of the interpreter.

The Interpretation and Use of Scores and Norms

Educational evaluation

Educational tests are a means of obtaining the basic data for evaluating the effectiveness of the instructional process and the school system. Only in terms of actual changes produced in students' behavior can schools justify their existence. This places a heavy burden on the testing procedures, since they must provide data concerning progress toward all objectives of education, not just the easily measured ones. The guiding principle in any study of test results of the students in a school system should be the improvement of the learning procedures so as to accelerate the rate at which the goals of education are being reached.

This educational accounting is a complex process and calls for all of the technical aid and assistance that test makers can supply. To evaluate the results of school groups, charts should be supplied on which can be plotted class and school results. The need for the types of comparability and meaningful units previously discussed are obvious here. Also it is clear that in order to evaluate the progress made by a specific group, many facts about them and the other groups being compared need to be known. These include such items as chronological age, scholastic aptitude, grade placement of subjects, amount of time devoted to the various subjects, and days in the school year. It is desirable that information be available concerning the mean values for schools varying in a known manner with respect to these variables. Distributions for the average scores for classes, schools, and school systems are needed, as well as those for individual students. Perhaps the greatest need for practical evaluation, however, is information as to the relation between course content and the requirements of practical daily life. Norms and standards for tests in these terms would seem to be of inestimable value to the administrator.

Educational guidance

In a society such as ours where assisting the individual to attain the optimal development of his potentialities is one of the fundamental goals of the group, it becomes of extreme importance to supply the individual with all possible knowledge concerning himself, his knowledge, his skills, his abilities as compared with those of others, and his special areas of strength and weakness. The importance of comparability for these purposes again appears striking. An individual profile on which the scores are plotted in terms of a common scale should be made available. Test scores in various fields should be grouped together so that the evidence concerning special areas of strength and weakness will be reinforced or

corrected. Arrangements should be made to enable the student to obtain a clear picture of the progress he has made during the previous year.

Many of the errors in interpreting test results are fundamentally to be attributed to inadequate or misleading information supplied by the test publisher for the proper use of test results. The confusion of norms with standards, the failure to show slow students that they are making real progress, and the unreasonable evaluation of instructional personnel or procedures would be much less prevalent if the types of information described in this chapter were available in usable form.

Selected References

1. Bergmann, Gustav, and Spence, Kenneth W. "The Logic of Psychophysical Measurement," *Psychological Review*, **51**: 1–24, 1944.
2. Brown, William, and Thomson, Godfrey H. *The Essentials of Mental Measurement.* Cambridge: Cambridge University Press, 1940.
3. Conrad, H. S. "Comparable Measures," *Encyclopedia of Educational Research*, ed. W. S. Monroe. New York: Macmillan Co., 1941. Pp. 340–44.
4. Courtis, S. A. "Maturation Units for the Measurement of Growth," *School and Society*, **30**: 683–90, 1929.
5. Cureton, E. E. "The Accomplishment Quotient Technic," *Journal of Experimental Education*, March 1937, pp. 315–26.
6. Flanagan, J. C. *A Bulletin Reporting the Basic Principles and Procedures Used in the Development of Their System of Scaled Scores.* New York: Cooperative Test Service of the American Council on Education, 1939. 41 pp.
7. ———. "Units and Norms in Educational Measurement," *National Projects in Educational Measurement*, ed. K. W. Vaughn. Washington: American Council on Education, 1947. Pp. 8–12.
8. Gardner, E. F. "The Determination of Units of Measurement Which Are Consistent with Inter and Intra Grade Differences in Ability." Ph.D. dissertation, Graduate School of Education, Harvard University, 1947.
9. Grossnickle, Louise T. "The Scaling of Test Scores by the Method of Paired Comparisons," *Psychometrika*, **7**: 43–64, 1942.
10. Kelley, Truman Lee. *Fundamentals of Statistics.* Cambridge, Mass.: Harvard University Press, 1947.
11. ———. "The Measurement of Overlapping," *Journal of Educational Psychology*, **10**: 458–61, 1919.
12. ———. "Ridge Route Norms," *Harvard Educational Review*, **10**: 309–14, 1940.
13. ———. "A Simplified Method of Using Scaled Data for Purposes of Testing," *School and Society*, **4**: 34, 71, 1916.
14. Kelley, Truman Lee; Ruch, G. M.; and Terman, L. M. *Manual for Standard Achievement Tests.* Yonkers-on-Hudson, N.Y.: World Book Co., 1925.
15 McCall, William A. *Measurement.* New York: Macmillan Co., 1939. 535 pp.
16. Mosier, C. I. "Psychophysics and Mental Test Theory: I, Fundamental Postulates and Elementary Theorems," *Psychological Review*, **47**: 355–66, 1940.
17. ———. "Psychophysics and Mental Test Theory: II, The Constant Process," *Psychological Review*, **48**: 235–49, 1941.
18. Rogers, D. C. "An Argument for Centile Ranks," *Journal of Educational Psychology*, **24**: 107–17, 1933.
19. Salvosa, Luis. "Tables of Pearson's Type III Functions," *Annals of Mathematical Statistics*, **1**: 1–125, 191–98, 1930.
20. Thorndike, E. L. "On Finding Equivalent Scores in Tests of Intelligence," *Journal of Applied Psychology*, **6**: 29–33, 1922.
21. ———, *et al. The Measurement of Intelligence.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. 616 pp.

22. Thurstone, L. L. "The Absolute Zero in Intelligence Measurement," *Psychological Review,* **35**: 175–97, 1928.
23. ———. "A Method of Scaling Psychological and Educational Tests," *Journal of Educational Psychology,* **16**: 433–51, 1925.
24. ———. "The Unit of Measurement in Educational Scales," *Journal of Educational Psychology,* **18**: 505–24, 1927.
25. Toops, H. A., and Symonds, P. M. "What Shall We Expect of the A.Q.?" *Journal of Educational Psychology,* **13**: 513–28, 1922.

18. Batteries and Profiles

By CHARLES I. MOSIER
Office of the Adjutant General, Department of the Army

COLLABORATORS: John C. Flanagan, *University of Pittsburgh;* Arthur E. Traxler, *Educational Records Bureau;* Alexander G. Wesman, *The Psychological Corporation*

IN ALMOST EVERY SITUATION IN WHICH PSYCHOLOGICAL MEASUREMENT is applied, more than one measurement is involved. The introduction of multiple measures of the same individual, or set of individuals, raises questions concerning how these several measures should be chosen, in what terms they should be expressed, and how they can be combined into a meaningful whole. The combination of the several scores for one individual may either merge the several parts or preserve their respective identities. This chapter is concerned with the need, the objectives, and the attributes of these sets of multiple measures and with the methods for combining them for interpretive purposes.

Definitions and Examples

The term "battery" is conventionally applied to a set of separate tests to be administered to the same group of individuals in order to meet a single measurement objective, or a closely interrelated set of such objectives. A "profile" is a graphic summary of the results of the multiple measurement which retains the identity of each measure until all are merged into a final action judgment. It should be recognized, however, that even though the identity of each measure is preserved up to the end, all related measures are ultimately combined in the decision to take (or recommend) action on the basis of the available data. Such a single action is involved whenever the set of objectives is sufficiently interrelated to justify considering the several measures as part of the same battery. It will be desirable, therefore, to consider the attributes of batteries in detail before turning to a treatment of ways of summarizing their results, including profiles as one such way.

Batteries of distinct tests

A consideration of the various kinds of batteries in common use will

bring additional understanding of the nature and purposes of multiple measurement and the ways in which it is given application. One of the most common notions of a battery is an assemblage of several tests, each constructed by a different author at a different time for a somewhat different purpose but brought together by the test user for his own immediate ends. The purposes of the several tests are compared with the various aspects of the objective at hand to select those tests likely to contribute to that over-all objective. Thus, a test of certain aspects of scholastic aptitude, for example, the American Council on Education Psychological Examination, might be combined with tests of achievement in the various major subject-matter areas to constitute an admissions battery for college entrance. The results of the several tests are likely, in the light of much previous research, to give a better prognosis of college achievement than is any one test alone. Again, tests of achievement in English, social studies, arithmetic, and science might be assembled from various sources to provide measures of school achievement. Such a battery is usually assembled *ad hoc* to meet some immediate, specific practical need. Though this is among the most common meanings of the term "battery," analysis of other examples will indicate that it is by no means the only one.

Some batteries, instead of being based on the *ad hoc* approach of the preceding paragraph, are prepared specifically to meet a more general integrated set of objectives. Thus, the ACE Psychological Examination is itself an assemblage of six tests, and yields not only a total score, but two "ability" scores, one diagnostic of aptitude in dealing with quantitative concepts, the other aptitude verbal in nature. Besides these two scores—Q and L—the six individual part scores may also be obtained for special purposes.

On a more elaborate scale, the Primary Mental Abilities battery provides in a single set of tests separate measures of a number of different and relatively uncorrelated abilities. Batteries of tests developed for the classification of personnel in the armed services, the General Aptitudes Test battery of the United States Employment Service, and the Differential Aptitude Tests are other examples of this approach. These batteries are distinguished from those first discussed by the fact that each part of the battery has been constructed as a part of an integrally conceived whole so that gaps and overlapping are less likely to occur than in assemblages of already existing tests.

In the area of achievement testing we find such batteries as the Stanford Achievement Test, the Iowa Every-Pupil Tests of Basic Skills, and the Metropolitan Achievement Tests. These batteries present an integrated set of tests in each of certain well-defined areas of instruction such as English, social studies, and arithmetic and are related to the graded cur-

riculums of the public school system. Moreover, the several components in the battery are, or purport to be, reduced to a single frame of reference for the reporting and interpretation of scores.

Omnibus tests: single scores

The two foregoing types of battery have been distinguished by physical segregation of the items into recognizable "tests," each with a beginning, an end, separately computed part scores, and (usually) a separate name. Strictly speaking, these batteries are the ones which meet the common definition. From the viewpoint of combination of measures, however, the essential characteristic of a battery involves multiple measures to achieve a single objective or a set of related objectives. In view of this conception, not all batteries are as immediately recognizable as are the foregoing. All of the essential characteristics of batteries are present in tests designated by a single name, with no physical segregation of the several types of items, and for which only a single score is computed. The scoring formula prescribes the method of combination and precludes further identification of part scores. These tests—for example, the Stanford-Binet, Henmon-Nelson, and other tests of the omnibus type—are in reality batteries where the components are (usually) not clearly identified. They consist of several sets of items, each set characterized by high intercorrelations within the set and relatively low correlations between any pair of items drawn from different sets. Some such "batteries," for example, the Wechsler-Bellevue Intelligence Scale, give recognition to the fact that several aspects of behavior are being predicted by deriving part scores even though the items comprising one part (Verbal or Performance) do not occur consecutively. Similarly the 1916 Stanford-Binet has been shown to measure a composite of several statistically independent abilities even though separate scores for each ability are usually not separately computed. In other areas, the Henmon-Nelson and Cooperative General Culture Tests may be viewed as batteries, to the extent that observations of several, relatively uncorrelated aspects of behavior are combined into a single measure. The methods of combining such observations will be considered later. At this point, however, Holzinger's conclusion (*13*) is appropriate:

> . . . A complete analysis of the data must involve as many averages as there are factors. [A test score is, of course, nx where n is the number of items and x the individual's average score on an item.]*
>
> The ordinary analyst might suppose that, even if several factors were involved in his data, a single average would somehow take them all into

* Bracketed inserts added by author.

account. While this is true to some extent, much important information may still be ignored.

From these considerations it is apparent that a single average [or a single test score]* as a complete summarization is justified only if the data are of rank one; that is, if only one common factor is involved.

Omnibus tests: multiple scores

A second sort of omnibus test, somewhat more readily recognizable as a battery, is exemplified by such tests as the Strong Vocational Interest Blank, the Bernreuter Personality Inventory, and the Kuder Preference Record. In this type of battery an aggregation of situations is presented in an order and arrangement which bears no relation to the personal attributes measured, but the component attributes (which may or may not be psychologically distinct or statistically uncorrelated) are later segregated by differential application of several scoring keys. Thus, although the items in the Bernreuter Personality Inventory are arranged in a scrambled order and numbered consecutively, the inventory is essentially a battery measuring four (or two or six) attributes—B1-N, B2-S, B3-1, and B4-D, or F-1-C, and F-2-C.

The single test as a battery

From these latter two examples, it is fairly easy to generalize the concept of a battery to include any test which is based upon multiple observations of the same individual—in other words, any test with more than one item. Thus, a test of 100 vocabulary items may be considered as a battery of 100 tests each consisting of one item and, if usual scoring formulas are applied, given nominal weights of 1 or 0. This extension of the concept makes it possible to apply to multiple-item tests the concepts and principles which have been developed for batteries, and to make the converse applications. In making these applications, of course, care must be taken to verify that the conditions of the test are fully accounted for and that the necessary assumptions are noted. This point is discussed more fully later in the chapter.

Measurement in relation to prediction

In the foregoing discussion, and elsewhere in this chapter, "prediction" is used in the sense of a statement made in advance of *knowledge of an event,* rather than in advance of the *event itself*. Thus, events occurring in the past, but presently unknown or imperfectly known, can be predicted in the same sense that future events can be predicted. The eruption of a nova which will not be visible in the heavens until 1980 is properly "predicted" even though the explosion on which it was based occurred

an astronomical number of years ago. Similarly the amount of algebra learned by Bill Jones last year may be "predicted" at any time until the amount of learning is actually known. In this sense, then, if in no other, all psychological measurement may be viewed as prediction, and the logic of prediction applied to give us a clearer understanding of what we do when we measure.

In another sense as well, we measure to predict, since it is not the individual's past behavior per se but his past behavior as an indication of what he will do or what should be done to him that usually interests us. We "measure" or grade an individual in algebra to know whether he will do well in analytic geometry or as an engineer. Will he succeed in the next course or in this vocation? Should he be required to repeat the course, or elect another major, or be denied a diploma? What level of training should be applied? What course of therapy is most appropriate? What is the prognosis? These forward-looking questions are implicit in all measurement, and the logic of prediction is applicable. Achievement test scores as descriptions of status alone are relatively sterile. They take on meaning only when they predict behavior—even if that behavior is simply better understanding of the English language in daily reading or listening, or fuller appreciation of the historical backgrounds of events which will occur five years from now.

In many practical applications of testing, particularly in the educational situation, this predictive use of measurement is remote (see chapter 16, pages 652–74). An end-of-course examination is generally viewed as "measuring what the student has learned." Prediction of what he will learn in a later course or of how he will apply his knowledge in a real-life situation is ignored in constructing, administering, and interpreting the tests. It is possible, of course, to divorce achievement testing from any predictive significance. To do so, however, not only robs the tests of their maximum usefulness and leads to inadequately conceived and constructed instruments, but obscures information which could be used for the improvement of teaching practices and curriculum organization. Much of what has been and will be said in this chapter relates measurement to its predictive function. The reader who is interested primarily in achievement testing and the measurement of the individual's current status should recognize that the discussion is equally applicable to his problems and that status measurement will gain new significance if its predictive implications are recognized and exploited.

The criterion as a battery

In practice, what we predict with a single test or a battery is usually

some criterion measure. When possible, this criterion is quantified, and statistical methods are used to make predictions. The criterion is not always quantified, and in some cases may not be directly observed. In such cases our predictions may be judgmental rather than statistical, or we may use some directly observable substitute (for example, total test score) instead of the criterion we are really interested in but cannot observe directly. Often in measuring school achievement, for example, any practicable "true" measure is unavailable. Teachers' ratings are probably not as valid as the test itself; it is not feasible to wait five years—or even one—to observe success in the next higher course (And how would it be measured?). Thus, the relation of the test items to the defined curriculum objectives, together with total score on the test itself, is taken as a basis for determining the extent to which each item predicts the ultimate, but unobservable, criterion. Nevertheless, in spite of our lack of a directly measured criterion, we still are using the test score to predict (that is, to estimate before the "true" answer is available) the student's achievement in the course.

Whether or not a measured criterion is available, and particularly when it is, the criterion is seldom simple. If measures of different aspects of the criterion are appropriate (and available), the criterion itself may be and should be considered as a battery (*5; 10; 16; 36; 32*).

When the criterion is considered as a complex of several measures of success, the various accepted aspects of success—quantity, quality, or time required for learning—present problems of combination just as do the several part scores of a battery. The methods of combination which are appropriate are governed by the same considerations as those applicable to the combination of tests. If, for example, the conditions of multiple regression are met (for example, that none of the intercorrelations among criterion variables approach unity) it is as legitimate to compute the multiple regression of a test on a set of "criteria" as the converse.[1] In some circumstances it is as proper to validate a company's hiring or promotional system, or a college grading system, against the results of a series of tests, as it is to validate the test *against the products of the system. The relationships between test and criterion are symmetric and reciprocal.* The various aspects of the criterion may be merged into a composite criterion until the identity of the several measures is ultimately lost, or, as we saw earlier, they may be kept as separate measures until the final action judgment. Thus, one may predict freshman honor-point average, in which perform-

[1] The fact that, in two sets of observations, time sequence or social considerations lead us to designate one as "test" and the other as "criterion" should not blind us to the fact that the criterion observations are themselves observations and subject to the same considerations of reliability and validity as is the test variable.

ance in each freshman course is combined (usually by arbitrary rather than statistical weights) into a composite criterion; or one may, on the other hand, predict course grades separately, combining them only in such decisions as "all predicted language grades are satisfactory; all predicted mathematics grades are unsatisfactory; this student should major in language and literature." In every case, however, there is some final judgment entered which represents a combination of the various items of information available. The choice here, as in all instances of multiple measures, is not whether to combine or not, but whether to combine statistically or intuitively (*5*).

Objectives of Multiple Measurement

The various types of multiple measurement represented by the commonly used varieties of batteries (including multiple-item tests) have now been reviewed, and it becomes pertinent to consider the three distinct purposes which lead us to obtain more than one measurement of the same individual.

The first of these is relatively unimportant, but nevertheless deserves mention here. A number of observations is sometimes used in order to provide finer degrees of discrimination among the individuals measured. This objective is particularly important when each observation is of an all-or-none character, as are most objective test items. If each item is scored only "right" or "wrong," then a ten-item test will yield at most eleven different scores. If more than eleven degrees of differentiation are required among the individuals tested, it is necessary (though not sufficient) to include more than ten all-or-none observations. If the individual observations were more reliable than they usually are, this purpose would assume greater practical significance than it actually does. In most situations, however, so many items are necessary to meet the second purpose to be described below that the need for finer differentiation is automatically provided. (How reliable the resulting differentiations are is another question.) If, for example, the tests in the Stanford-Binet scale were so reliable and so accurately placed that each individual tested passed all items up to the base year, failed one or more items in the next mental age group, and failed all tests beyond that point, the provision of only six to eight "levels" among all those having the same basic mental age—for example, twelve years—might give too few discriminations. Since such is not the case, however, there seems to have been little reason for complaint that the scale fails to discriminate among those with the same "base year."

The second, and one of the principal objectives, of multiple measurement is to increase the reliability of measurement. Since each observation

has some elements of unreliability, the reliability of a composite of repeated measurements of the same fundamental attribute is increased by increasing the number of observations (see chapter 15, pages 580–81). This conclusion does not hold, of course, if the added measurements have substantially greater error variance than the measurements already at hand. (Whether this increase in reliability will follow the Spearman-Brown formula will, of course, depend on whether or not the basic assumptions regarding the relationship between the additional measures and the original measures are fulfilled.) It should be noted here that when increased reliability of the final measure is the objective of multiple measurement, the several measures should all be highly intercorrelated, since all are intended to be measures of the same fundamental characteristic.

In a modification of the same objective, a battery of tests is sometimes used to provide more reliable measurement of a single common factor. In this modification, the objective of the average is to minimize the effects, not only of the randomly distributed chance error, but of the systematic effects of a large number of uncorrelated specific factors as well. Thus Binet, in his earliest measures of intelligence, summed the score of the individual on a number of widely different tests. All of the tests were selected with the idea that each one measured some factor common to the group of tests. Each one might also depend in part on some other factor as well, but this factor was (hopefully) not shared to any marked extent by any of the other tests. In this instance, the specific factor of each test was measured reliably and hence could not be considered as variable or chance error. Even so, the contribution of any one of the specific factors (and since they were uncorrelated with each other, of the aggregate) was small in relation to the contribution of the supposed common factor of intelligence.[2] If the postulated conditions are met, we may expect that the several tests will have such high intercorrelations that, when corrected for specific and error factors, they are all unity.

The third objective of multiple measurement is to provide measures of *unrelated* aspects of the behavior-to-be-predicted. ("Unrelated" is used here in the sense of "uncorrelated in the population of individuals.") This is the objective in mind when tests designed to measure various aspects of aptitude (verbal, spatial, perceptual), achievement in specific subject matter (mathematics, English, history), and motivation (interests, study habits, etc.) are combined into a battery for the prediction of scholastic success. Included in this objective is the use of suppression variables (to be considered in more detail later). In this use, the purpose is not to secure

[2] For the extent to which the hypothesis was met see Wright (*40*). Also see McNemar (*19*) for a discussion of the evidence for the present versions (Forms L and M).

repeated measures of the same fundamental aspect of behavior in order to reduce the effect of specific factors and errors of measurement; rather it is to measure *different* aspects of the criterion, each related to that criterion, but not duplicating each other. In general, then, in the selection of measures to be included in a battery designed to serve this objective, we seek those which show substantial correlation with the behavior to be predicted, and absence of correlation with one another. Exceptions will be noted to this general rule, and the terms "absence of correlation" and "lack of correlation" will require more precise definition.

A more complicated variant of the third objective of multiple measurement is that of multiple prediction—the gathering of information which will predict simultaneously for several criteria. This is frequently the objective of guidance or placement testing, where we seek to predict in which courses, vocations, or jobs the individual is most likely to succeed. Thus, a uniform battery of tests may be administered to all entering freshmen in order to guide them into the proper courses or curriculums. The tests will be differently weighted (some weighted zero), in all probability, for the several curriculums considered, but each test should be a useful predictor for more than one curriculum. This variant is subject to the same principles as simpler instances of the third objective (*41; 42*).

Since objectives two and three have often been confused, particularly in connection with multiple-item tests, it will be worth while to recapitulate the differences between them. Let us consider a criterion R (the behavior to be predicted) and a set of tests (or test items) $A, B, C. \ldots$ If the criterion is simple, so that it is some function of the single trait X, then we seek tests $A, B, C, \ldots$ such that each test is also some function of X and of any factors unique to the particular test. We may symbolize this as follows:

$$R = F_1(X)$$
$$A = F_2(X, U_A)$$
$$B = F_3(X, U_B)$$
$$C = F_4(X, U_C)$$

Since it is to be hoped that the contribution of X to the variance of A, B, C will be large in relation to the contribution of the unique factors U_A, U_B, U_C, it is to be expected that the intercorrelations of $A, B, C, \ldots$ will be large. Essentially we are here combining tests $A, B,$ and C to give us a more accurate measure of the single underlying variable X—and thus are concerned with the second objective. One concrete example of this is the measurement of knowledge of English usage (symbolized by R above). It is presumed that there is here a common factor X to be measured and

that this factor is reflected in tests of A (Grammar), B (Punctuation), C (Word Usage), etc. Each such test—for example, Grammar measures not only the common factor X, but also measures a number of factors (U_A) not measured by either of the other tests, Punctuation and Word Usage. Thus, the sum of scores on the three tests will be much less affected by these specific factors, and much more representative of X, than will the score on any one of the three tests.

If the criterion is complex, so that it is a function of several different aspects of behavior, or traits, say X, Y, Z, then we seek tests A, B, C, so that each aspect of the criterion will be included, and we also seek efficiency in our testing, so that no aspect will be duplicated.[3] This situation can be symbolized as:

$$R = F_1(X, Y, Z)$$
$$A = F_1(X, U_A)$$
$$B = F_2(Y, U_B)$$
$$C = F_3(Z, U_C)$$

To the extent that each of the tests, A, B, C, meets the condition that it depends on a separate factor, X, Y, Z, and on factors unique to that test, the intercorrelations of ABC will be zero. As an instance of this situation we may consider an example from another field of measurement. (Please note that the choice of examples does *not* imply that the examples are characteristic. Either of the two objectives may be found in any field of measurement.) Let us suppose that we are attempting to predict success as a machinist (symbolized by R above). Let us suppose, moreover, that success as a machinist involves certain aspects of shop arithmetic, finger dexterity, and knowledge of tool usage (X, Y, and Z respectively). Our three tests, A, B, and C, each depend on the pertinent aspect of its respective knowledge skill or aptitude and on extraneous considerations (specifics and chance error) as well. It is clear that we use three tests here for different reasons than those which led us to use three tests to predict knowledge of English usage.

The use of suppression variables (*14*, pp. 430–47) can also be presented in a similar scheme. Let us suppose that scores in test A reflect ability in factor X which is related to the criterion, and in another factor, W, which is not. Then the effectiveness of A in predicting R is diminished by A's partial dependence on W. If we can introduce a new test, D, for which

[3] We may, in order to increase the reliability of one test, for example, Test A, wish to include another measure, A', of the same attributes. If we do, we may not include both A and A' separately in the battery for which multiple regression weights are to be computed. The proof of this is presented on page 787.

scores depend on W but not on X, then D will correlate somewhat with A (because of the common factor W) but not with the criterion R (since W is not common to R). Despite the zero correlation of test D with the criterion, D will contribute to the effectiveness of the total battery, $ABCD$, since its effect when properly weighted will be to remove from A the distortion which results from A's partial dependence on W.

Reliability of Multiple Measurement

Although the emphasis which has been given in the literature to the problem of reliability far outweighs its importance, particularly in relation to the importance of predictive value, it is still a concept which necessarily pervades the entire field of measurement. The general concept of reliability of test scores is adequately covered in chapter 15. Only those aspects directly related to the use of batteries and profiles will be touched upon here, and those but lightly.

Reliability of components of the battery

Obviously, if any confidence is to be placed in the results of administering a battery of tests, it is essential that each component of the battery possess some degree of reliability if its inclusion is to make any contribution. (This is, of course, a necessary but not a sufficient condition to its contributing to the objective of the battery. The test may be reliable and contribute nothing; if it is unreliable it *cannot* contribute anything.) How great a degree of reliability is required will depend upon the purposes to which the battery is to be put, on the methods for combining scores or utilizing the results, and on the interrelationships among the battery components. In many cases, the scores on an individual test in the battery are to be interpreted singly, as well as in combination with the other components. In such instances that test must be just as reliable as if no other were to be included in the battery. This is the case in many diagnostic tests where part scores are obtained and interpreted as indicative of particular weakness at the same time that an over-all score is obtained and used for other purposes. It is also true of school achievement batteries if the pupil's standing in any particular subject is to be evaluated in addition to his over-all standing. Such part scores or individual test scores can be used with confidence only if the reliability of the particular part warrants reliance. While this may seem to be a truism, it is one "more honored in the breach than in the observance." The fact that the reliability of the total score is satisfactory cannot legitimately be (but occasionally is) adduced as a basis for interpretation of its various parts.

Moreover, as will be developed more fully later, if interpretation is to be based, not only upon the scores on each particular part, but also upon

differences between those scores, then the requirements for component reliability become still more stringent, and the intercorrelations among the parts must be examined as well. Much of the impressionistic interpretation of batteries (including inspection of score profiles) depends on such statements as, "John scored higher in mechanical ability than he did in clerical ability [or, higher in English than in mathematics]; therefore. . . ." Thus, the reliability of test score differences is at least as important as the reliability of test scores. It suffices at this point to state that for such interpretation it is important that the reliabilities of the two tests being compared must be high, and equally important that the correlation between them be considerably lower than their reliabilities.[4]

Reliability of composite derived scores

The reliability of a weighted sum of elements, when the reliability of each is known, may or may not exceed the average reliability of the component elements. The general formula for estimating the reliability of such a weighted sum is given by:

$$r_{11} = 1 - \frac{\sum_{j}^{n} w_j^2 - \sum_{j}^{n} w_j^2 r_{jj}}{\sum w_j^2 + 2\sum_{j}^{n}\sum_{k}^{n} w_j w_k r_{jk}}, \tag{1}$$

where $j < k$.

This formula is applicable regardless of whether or not the weights, w_j, are obtained by a least-squares procedure or by intuition. This reliability of a composite score needs to be differentiated sharply from the multiple correlation between the composite and the variable to be predicted. An example may serve to clarify the distinction. Let us suppose that a battery of three tests, Mechanical Information, Block Counting, and Surface Development, is used to predict instructor's grades in shop mechanics. (We might equally well have supposed tests of French Grammar, Reading Comprehension, and Vocabulary to predict grades in second-year French or to "measure attainment" in first-year French.) Let us suppose, moreover, that the battery has been given twice, but scores from only the first administration are used to predict success. The multiple correlation coefficient, $R_{1.234}$, is the correlation between the best (least squares) weighted combination of the test scores and instructor's grades. If, however, the grades were to be predicted from the second administration, using the same regression

[4] See discussions in Segel (*29*) and Bennett and Doppelt (*3*). For further discussion of the problem of differential prediction, see also Thorndike (*42*) and Brogden (*41*).

weights, the correlation between first prediction and second prediction would represent the reliability of the weighted composite. There is no need to point out here that the multiple correlation is the acceptable measure of the effectiveness of the battery in predicting grades. Moreover, the formula above shows that the reliability of the weighted composite is dependent upon the particular set of weights used, just as the reliability coefficient of a single, multiple-item test is dependent on the scoring formula.

It is pertinent to note that when the intercorrelations of the components are all zero (an unlikely circumstance), formula (1) reduces to:

$$r_{11} = \frac{\sum_{j}^{n} w_j^2 r_{jj}}{\sum_{j}^{n} w_j^2}$$

so that the reliability of the weighted composite is the weighted *average* of the component reliabilities, a value usually less than that for the highest component reliability. The tests in a battery used to measure various aspects of whatever is being measured are, however, often selected to have intercorrelations as low as possible. It can be readily seen that in such circumstances a combination of short, relatively unreliable tests will not yield increased reliability (but rather less) for the composite scores, whether the composite be some simple summation or a least-squares prediction of an outside criterion. Improved prediction (or measurement) can be achieved by increased reliability of the measures as well as by increasing the number of tests in the prediction battery.

The notion of increasing the reliability of a test (and hence its maximum validity for some purpose) by increasing the total number of items is well established by a tradition of uncritical acceptance. Therefore, an additional warning may be appropriate here where the combination of tests into a battery results in an increase in total number of items. As has been pointed out, the Spearman-Brown formula for estimation of the reliability of a test "n" times as long as one of unit length holds only if each test of unit length has the same variance and reliability as every other such test in the composite, and if, moreover, the intercorrelations of the "unit tests" are all unity when corrected for attenuation. There are few prediction batteries in which these conditions are likely to be met. In fact, the last condition is incompatible with the method of multiple regression.[5]

Another problem arises in connection with the reliability of the tests included in a battery. This problem is most important when the scores on the individual tests are retained in uncombined form and interpreta-

[5] Compare with footnote 3, p. 773.

tions are based on the relations between the scores on the various tests.[6] Thus in such batteries as the Strong Vocational Interest Blank or the Kuder Preference Record, for example, the counselor has before him scores on a number of different occupational or interest areas. His interpretation of those data may be direct, but it is frequently differential; that is, he may conclude that this subject has a high (or a low) degree of interest as an accountant, but he will also likely infer that this subject has a stronger degree of interest as an accountant than he has as a lawyer. The school counselor is equally concerned with whether the pupil is stronger in science than in social studies. It is with the second type of interpretation that we are here concerned. Whether the comparison is between two or among several scores, we are dealing, in the last analysis, with differences between test scores. The reliability of that difference must, therefore, be a matter of concern. The reliability of the difference of two fallible test scores is obtainable from appropriate modification of the derivation leading to formula (1) and is seen to be:

$$r_{tt} = \frac{r_{11} + r_{22} - 2r_{12}}{2(1 - r_{12})}, \tag{2}$$

or, if the two tests have identical reliabilities,

$$r_{tt} = \frac{r_{11} - r_{12}}{1 - r_{12}}. \tag{3}$$

The effects of this relation can best be seen by assuming hypothetical values for the test reliabilities and for the intertest correlation. These are shown in Table 14 for the difference between two tests, each with the same reliability as the correlation between them varies from .80 to .00.[7]

TABLE 14

RELIABILITY OF DIFFERENCE SCORES IN TERMS OF THE CORRELATION BETWEEN THE SCORES AND MEAN RELIABILITY OF THE SCORES*

Mean Reliability	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
.90	.90	.89	.87	.86	.83	.80	.75	.67	.50	0.00
.80	.80	.78	.75	.71	.67	.60	.50	.33	0.00	
.70	.70	.67	.62	.57	.50	.40	.25	0.00		
.60	.60	.56	.50	.43	.33	.20	0.00			
.50	.50	.44	.38	.29	.17	0.00				
.40	.40	.33	.25	.14	0.00					
.30	.30	.22	.13	0.00						
.20	.20	.11	0.00							
.10	.10	0.00								
0.00	0.00									

* This table was prepared by and is reproduced through the courtesy of Edmund Fuchs and Mrs. Bertha M. Harper.

[6] Cf. pp. 774–75.

[7] See also Bennett and Doppelt (3) for an abac setting forth these relationships.

Obviously, in most practical testing situations, particularly in achievement testing, but in aptitude testing as well, zero correlations between tests are unlikely. Achievement in two school subjects depends on so many overlapping factors that sound measures of each will necessarily correlate to some extent—usually with coefficients ranging from .40 to .60. This fact (to the extent that the correlation is based on true overlap between the school subjects, and not on artifacts of testing) does not reduce the effectiveness of the tests considered singly. It does, however, call for caution in interpreting differences between them whether presented numerically or in a profile. It also imposes an additional obligation on the constructors of tests to be so used to make the individual tests so reliable that *differences* will be reliable despite the correlations among the tests.

As an example of this, the authors of the Iowa Tests of Educational Development made it their objective in constructing the tests that all tests in the battery be uniformly reliable at the level necessary to provide an average coefficient of reliability among the difference scores of .80. Moreover, to aid interpretation and to prevent unwarranted reliance on insubstantial differences between scores, the scaled score unit has been taken at one probable error of measurement. Thus, for a difference of *one* scaled score point between John and Mary on the English test (or between the scores of two schools) the probability is even that a second testing would result in a difference in the same direction.

Attempts such as this to render tests foolproof in interpretation as well as in administration have not received the attention they deserve. Some work has been done, and each of the widely used batteries has made its own contribution toward methods of interpretation which minimize the possibility of misinterpretation by test users who are not expected to be as familiar with the statistical pitfalls of test construction as is the test constructor. Much more work needs to be done, however, to provide interpreted scores which are not open to fallacious conclusions. Moreover, until the millennium when all tests will be constructed in this manner, each test user has the obligation of any craftsman to know not only the uses, but also the limitations, of the tools he uses.

The problem of interpreting differences between test scores and the interrelationships among a set of scores will be considered further under profiles.

Combination of Multiple Measurements

Inevitability of combining scores and of weighting

In almost every conceivable situation in which a test score is obtained for an individual, the question of combination of measures enters the

picture. Let us take as our simplest example the administration of only one test consisting of only one item to a single individual. There will be few situations indeed in which that single test score will be interpreted as an isolated fact. Rather, someone—teacher, clinical psychologist, vocational counselor, or appointing authority—will consider that test score in relation to other known facts about the individual—his age, his grade in school, his expressed vocational choice, the impression he creates in an interview. In the instance we håve cited, the combination of data is informal, non-statistical, and wholly judgmental. Nevertheless, in arriving at the over-all judgment leading to action—"pass or fail," "consider some other vocation," "hire or reject," "recommend therapy"—the several facts to be considered in connection with the single test scores are "combined" with it. Moreover, each fact is weighted in that estimate in accordance with the interpreter's judgment of its relative importance in the total complex. This combination and weighting occurs nonetheless inevitably because it is informal and judgmental. Not only is some method of combination almost inevitable, but there is also the possibility that injudicious weighting of elements in a judgmental combination may nullify the effectiveness of the most significant measurement.[8] It is apparent, then, that the various techniques of combining measures should be given even more importance than they have already assumed, particularly in relation to the interpretation of test results.

The foregoing demonstration that the use of multiple measures without any attempt at statistical combination leads almost necessarily to judgmental combination does not imply that statistical combination is the *only* defensible technique. There are numerous situations where the factors to be taken into account in reaching conclusions from a complex set of data are too many and too vaguely defined to be amenable to statistical treatment. *Statistical methods for combining scores are not substitutes for, but aids to careful thinking.* Nevertheless, if we are to aspire to a scientific study of individual differences, the intuitive judgments of the counselor, clinician, or other interpreter of test data must be brought within the framework of scientific method. Whether the interpretation of a set of test scores is based on least-squares regression weights or on the clinical experience of the psychologist, the interpretations must be submitted to the basic tests of reproducibility by another competent investigator and of demonstrated relationship to other, significant, phenomena, that is, to the familiar

[8] During the war a number of studies showed that the validity of prediction from test scores evaluated in an interview (and based in part on the additional information developed in the interview) was usually somewhat lower than the validity of the statistically weighted test scores alone. A more recent study confirming this is the follow-up of VA clinical trainees as reported by Kelley and Fiske (*18*).

tests of reliability and validity. The studies thus far made indicate that statistical methods of combination, while far from perfect, have a distinct advantage in most situations over the informed judgment of a single individual.

Wherever statistical combination is possible, the test scores should be so combined to give the best prediction of the criterion. If judgmental interpretation is necessary to take other, nonstatistical data, into account, all statistical data should be combined into a single best prediction of the criterion, and this composite then combined with the non-quantitative data. In this way judgmental errors affect only the combining of the two types of data and are precluded from entering into the combination of test scores, whatever error may remain in the over-all judgment. As an example, several tests and a personal interview may be the basis for admission to college. In this situation the interview results should be judged in relation to grade-point average predicted from the composite tests, rather than in relation to the several test scores. In this way, the number of variables to be combined judgmentally is reduced to two—predicted grade-point average and interview appraisal. It remains to be demonstrated in any particular case that the judgmental combination of interview with grade-point average predicted from tests will add to the prediction based on test scores alone.

There is another aspect of the application of judgment in the combination of measures. In many situations where a measurable criterion of success is present, that criterion is incomplete. For example, we may have course grades as the only measured criterion of school success. Nevertheless, the stated objectives of the school may include the development of skills in social adjustment. If we investigate the relative effectiveness of a scholastic aptitude test and a measure of social adaptability in the prediction of the measured criterion (grades), it is likely that the aptitude test will receive great weight, the measure of social adaptability little or none. In this case it may be argued by some that the relative weights are artifacts of the defects in the criterion and that our judgment of the "true" criterion requires modification of them. While such judgmental modification may be justified, it should be undertaken, if at all, with extreme caution, by a highly competent judge and with full realization that while it may better the weights, it may also worsen them materially.

In the situation of a multiple-item test, where the pattern of responses to the entire set of items is to be combined in a single test score, the argument of the preceding paragraph is even more applicable. Scoring formulas are many and varied. Each one, though, represents a method of combining multiple measures into a single composite and must be analyzed as such, rather than merely as a way of "defining" the test score. Most formulas are alike in obtaining for each individual a score on each item, and in

assigning a weight to it. The composite score for an individual is the sum of the products, item weight times item score.

Moreover in most tests all items are assigned identical nominal weights of unity, and the item score takes only the two values, 1 if "correct" and 0 if "wrong." These are used for reasons of expediency and should not blind us to the two components (weight and score) which are involved. As a matter of fact, other methods are possible, as we shall see later. Two considerations, however, are taken up here.

The first is the relation between nominal weights and real weights, thoroughly explored by Richardson (*26*). It is not possible to recapitulate Richardson's analysis here. The conclusion of principal relevance is that even though the apparent weights are equal, the effective weights of the individual items (or part scores) are usually not. Instead, each is proportional to the variance of the item or test (pq or σ^2) and to the sum of its correlations with the other items or tests in the set, $\sum_{j}^{n} r_{ij}$.

The second conclusion of utmost significance is that set forth by Holzinger (*13*). He has shown that a score based on the sum (or average) of a set of scores for a single individual is a measure of that individual's score on the first centroid factor of the set of tests. This is a psychologically meaningful score only if all of the tests measure a single common factor, that is, if the purpose of the set of tests, or test items, was to increase the accuracy with which a single aspect of behavior is to be measured. If the set of tests or items measures more than one aspect of behavior, then some other type of combination than simple, unweighted averages must be utilized for the most meaningful result since the *centroid* factor is unlikely to be the most predictive composite from the set. (Note that an unweighted average is not one in which all items are unweighted, but one in which they all have the same nominal weight.)

Except for simple summation or averaging, the statistical method of combining scores in most common use is that of multiple regression. The assumption most frequently made is that the behavior to be predicted is proportional to some weighted sum of the scores on the several tests. The two problems of such linear multiple regression are (1) to determine the weights to be applied to each test so that the total squared discrepancy between predicted behavior and the actual observed behavior for a group of individuals will be as small as possible; and (2) to estimate the accuracy of prediction through some such measure as the standard error of estimate or the multiple correlation coefficient. Multiple regression as a method of combination is discussed more fully at a later point. There are other methods which should be considered first.

We have already considered the judgmental evaluation by the interviewer

(teacher, psychologist, or counselor) as a special case of weighting the several factors (test scores) by a set of weights determined subjectively by the interviewer rather than by statistical summarization of experience.

Test scoring as a problem of multiple measurement

The foregoing discussion has indicated in general terms several of the more widely used methods of combining the results of a number of measures into a single composite score. A number of these are applicable to test scores as usually defined. Items of a test are usually combined by a simple count of the number right or by a scoring formula which introduces a correction for the number "answered right by chance." In any event the formula is:

$$S_i = \sum_{j}^{n} w_j x_{ij},$$

where S_i is the composite score of individual i,
w_j is the weight given the j-th item, and
x_{ij} is the item score for individual i on item j.

We may limit the values of x_{ij} to 1 ("right") and 0 for all other responses; we may extend the range of values to 1 for right, 0 for omission, and $\frac{-1}{k-1}$ (where k is number of choices) for recorded wrong answers; or we may assign a separate value of x_{ij} to each possible item response. We may assign all items the same weight or combine weight and individual's score into tabled products so that the composite is recorded (on the scoring key) for each possible differentiable response x_{ij}. Whatever mechanics we use to simplify our scoring labor, we have still the combined problem of defining the allowable item-response values and of determining the values of w_j which are optimal for the purpose at hand.

There is often no external criterion available; in other situations the nature of the problem under investigation makes questions about the nature of the "traits" defined by the tests more significant than an external criterion. In these situations, data supplied from the test battery itself become the basis for defining the nature of the combination. Lest there be those who believe that an internal criterion is used only because no external one is available, may we point out that a systematic investigation of traits and a determination of basic mental abilities through the techniques of multiple factor analysis are situations in which an internal criterion is used because it is best, rather than because it is expedient.

Thomson (*32*) and Mosier (*24*) have pointed out ways of weighting the parts of a battery so that the resulting composite has the maximum re-

liability. Even if measurement is undertaken in the absence of any explicit prediction, there is always the implicit prediction: "What would the individual's score be on a 'true' as opposed to the 'fallible' measure of success?" Thus, this method is primarily useful when all of the components are known to be measures of the same basic attribute and more than one measure is used to give a more dependable measure of that attribute; more accurate prediction is then possible from this more dependable measure.

Although this is not the place to discuss the techniques of item analysis, it should be pointed out in passing that the use of these procedures implies combination of measures and thus weighting as well. In a simple example, an experimental test of 200 items is reduced to a final test of 150 items through the application of item analysis techniques. In a real sense the 50 discarded items have been assigned weights of zero, while the 150 retained items are assigned weights other than zero, usually unity. The various techniques of item analysis can be grouped into those which yield item indices proportional to regression coefficients b_{yx}, or b_{xy}, or to correlation coefficients. The techniques of Toops (*34*), and of Wherry and Gaylord (*39*) provide item indices which are functions of the item's contribution to multiple regression. If the item analysis is in terms of total test score as the criterion, it becomes a method of maximizing the reliability of the test for a given length of test. The method of reciprocal averages extends the technique of weighting for maximum internal consistency to a weighting (or differential scoring) of the individual responses to maximize the separation among individuals (*15; 23; 21; 22*). Empirical research is needed to determine whether in this method the determination of a set of optimum weights will converge to a single, principal axis solution, or can be extended to provide, through appropriate selection of the initial weights, optimum weights for each primary factor included in the battery.

Present techniques for analyzing the interrelations of the components of a battery of tests into the underlying factors include methods for the estimation of factor scores. These weighting procedures constitute another instance of combination of measures in terms of internal, rather than external, criteria. The General Aptitudes Test battery developed by the United States Employment Service, Thurstone's Primary Mental Abilities battery, and the Differential Aptitudes battery developed by the War Department for use in the program of separation counseling, represent this type of approach. Obviously, for practical application, the effectiveness of such derived "factor scores" depends on relating the small number of "factor scores" (instead of the larger number of test scores from which they were derived) to the socially significant behavior which it is desired to predict, and on developing tests which truly and completely measure the factors

they are intended to represent. If the tests do not measure completely, they are lacking in economy; if they are unrelated to external criteria, they are of little or no use.

Prediction of an external criterion

The most common procedures used for the combination of the scores obtained from the individual tests included in a battery relate to techniques for the prediction of some outside criterion, either explicitly or implicitly stated, and related directly or indirectly to the behavior to be predicted. Examples of explicit statements of the criterion will occur to every reader. Prediction of school grades, job success, marital compatibility, and psychiatric diagnosis are common instances. The concept of the mental age is an instance of an implicit criterion (chronological age predicted from test performance) which is indirectly related to the ultimate criterion—behaving intelligently in certain types of problem situations. Thurstone (*33*), Richardson (*27*), and others have pointed out the fallacies involved in the application of the mental age concept. The analysis of the criterion and its measurement in reliable *and valid* terms is one of the most challenging problems faced by mental testing.

The most frequently used approach to the combination of measures for the prediction of an outside criterion is that of multiple linear regression. As we have seen, the basic assumption underlying this approach (and other similar approaches) is by no means the only one possible, even though it is one of the simplest. It may well be that many forms of behavior can be adequately predicted by a simple additive combination of test scores; and when this is the case, there is no justification for introducing more complex modes of combination. Nevertheless, when the prediction from a *linear* assumption is not as close as might be desired, the possibility of other hypotheses should not be overlooked.

The assumption that the criterion behavior is a linear sum of the test scores is only one of a large number of assumptions which might be made. The fact that linear combination is the simplest assumption should not blind us to the existence and possible superiority of other methods of combination. The possibility of such other hypotheses has long since been pointed out and only occasionally investigated. The relative value of linear and nonlinear combinations should be more thoroughly investigated before the linear hypothesis is uncritically accepted for each set of scores. It should be pointed out that where nonlinear relations are obtained, they frequently do not stand up under cross validation, and even when cross-validated many more cases are needed to determine stable relations for nonlinear than for linear combinations.

Certain corollaries of the assumption of linear dependence make its universal applicability questionable. If a certain type of behavior is a linear function of two abilities, then however low an individual may be in one of the abilities, he can reach "satisfactory" performance if only he compensates for his deficiency in the one by scoring sufficiently high in the other. A second such corollary is that for a constant score in one "ability" there is a corresponding increase in the criterion for each increment in the second ability and that this increase holds true *throughout the total range* of the second ability. The inapplicability of both of these corollaries to certain situations may be seen in the example of the relationship of visual acuity to skill in driving a car. Regardless of how high an individual may rate in all the other aspects which contribute to driving skill, if his visual acuity is below some lower limit (and that probably well above total blindness), his skill will never reach an acceptable level of safe driving. Moreover, for the same level of ability in the other components of driving skill, any increases in visual acuity beyond some other critical value (probably well below 20/20) will not result in any observable increase in driving skill. Figure 69 will illustrate the hypothetical relationship between driving

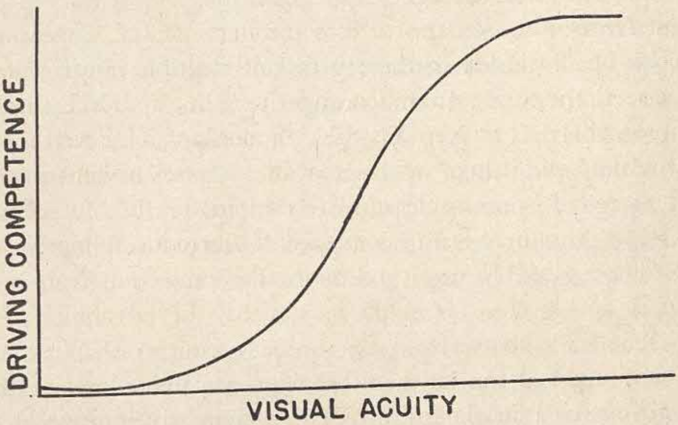


FIG. 69.—Diagram illustrating the hypothetical relationship between driving skill and visual acuity over the entire range.

skill (behavior to be predicted) and visual acuity for constant ability in all other components. It may well be that the central portion of the figure is reasonably well fitted by a straight line. For that range of abilities the linear hypothesis is a satisfactory substitute for a more exact statement of the relationship. Since this middle range (or, in general, some narrow range) is sufficient to cover the range of "normal" persons, linear hypotheses often serve satisfactorily for such normal individuals. The clinical psychologist has sought qualitative interpretations and has tended to reject

quantitative ones. It may be that one reason for this has been the quantitative psychologist's attempt to use only the simplest (linear) hypothesis and that his hypothesis, though satisfactory for the normal range, breaks down in the area of deviate behavior with which the clinician is concerned. In any event, the psychometrician should not, in the interests of simplicity, overlook nonlinear hypotheses when these are called for by the data or by rational considerations. At the same time, a word of caution should be added; in many extensive investigations, particularly in the armed forces during the war, nonlinear relations between test and criterion were sought and seldom found.

Simple linear correlational algebra, however, will simplify much of the clinician's or the school counselor's labor. Rapaport (*25*) computes for every individual the standard score difference between each possible pair of tests in the battery. He uses the N difference-scores thus obtained, and computes the means and standard deviations of the difference scores for each diagnostic category. These same means and standard deviations can, of course, be obtained simply from the means, sigmas, and intercorrelations of the individual tests. In this connection, Brogden (*4*) has proposed that where linear hypotheses are inadequate, new test variables be defined in terms of the squares or cross products of test scores and these new variables be included in the regression equation. Thus, if we wish to predict a criterion, X_0, from scores on tests X_1 and X_2, we may add three new variables $X_3 = X_1^2$, $X_4 = X_2^2$, and $X_5 = X_1X_2$ and investigate the effect of including these in the regression equation. This method has the disadvantage of using up degrees of freedom rapidly. Moreover, if it is employed, the cautions set forth on page 787 must be observed, a large number of cases should be used, and the findings cross-validated in another sample.

The use of multiple regression techniques assumes that the criterion score of an individual can be predicted from a weighted sum of the predictor tests included in the battery. The weights to be applied to each variable are determined in such a way that, for the sample on which the weights are determined, the total squared discrepancy between predicted and obtained criterion scores is as small as possible. The method provides also, in the multiple correlation coefficient, a measure of how closely the obtained scores are approximated by the predicted scores. Note that the weights capitalize on any sampling fluctuations in the sample and should thus be verified on another sample.

If, and only if, the several predictor scores and the criterion are normally distributed and the regression is linear, then the average discrepancy (predicted-obtained) will be the same throughout the range. If one or more

of the variables is not normally distributed, then, as in the case of simple regression, the standard error of estimate may vary from one part of the range of predicted scores to another, and the value computed from the relationship $\sigma_{0.123} = \sigma_0\sqrt{1 - R_{0.123}^2}$ is merely an average of values which vary throughout the range of predictors.

In the application of multiple regression techniques to any particular battery, care should be taken that no two of the predictor variables have correlations approaching unity. In view of the measurement objective for which multiple regression is an appropriate technique, this situation should not arise (compare with pages 770–72). If, because of misapplication of the technique, it should happen that two tests in the prediction battery do show correlations approaching unity, the multiple regression weights for all variables and, consequently, the multiple correlation coefficient approach indeterminancy. The reason for this can be readily seen if the normal equations are written out and the solution for each of the regression constants in turn expressed in determinantal form.

$$\beta_{01.23} = \frac{\Delta_1}{\Delta},$$

where Δ is the determinant of intercorrelations among the predictors and Δ_1 is the determinant formed by replacing row 1 and column 1 of Δ by the correlations with the criterion.

If any two variables in the determinant of coefficients, say variables 1 and 2, have correlations approaching unity, then we may expect logically that the correlation of variable 2 with each of the other variables in turn will approach very closely the correlation of test 1 with the same variable. When this occurs, the entries in the two columns (or rows) 1 and 2 will be proportional, or nearly so, and the value for the determinant of coefficients which represents the denominator of the fraction on the right of the equation for each regression weight in turn will be zero. Similarly, for every regression weight except those for variables 1 and 2, two columns of the determinant in the numerator will also be proportional, and for all weights except w_1 and w_2 the right-hand side of the equation reduces to the indeterminancy, 0/0. This is merely the simplest case. The same conclusion holds when any test in the battery can be wholly predicted from any *set* of the remaining predictors. If we were dealing with exact values freed of chance errors and with correlations of exactly unity, this indeterminancy would show up in the routine application of the numerical solution for beta weights. Since, however, in ordinary problems, we deal with less than ideal conditions, the essential indeterminancy may be masked, and only care in the

formulation of the problem, and a verification from the nature of the results that the multiple regression technique is appropriate, will save the investigator from reporting nonsensical results in all good faith.

In the selection of predictor variables, emphasis is frequently placed on the correlation of each predictor with the criterion, and the effect of the correlations among the predictors is slighted. We have seen above one instance of how such oversight can lead to nonsensical results. In other instances the emphasis on high criterion correlations can result in less than the maximum effectiveness of prediction. It is quite possible, and frequently feasible, to include in the battery a test which, though having a correlation of zero with the criterion, will nevertheless increase the multiple correlation of the battery of which it is a part (*14; 37*). The "clearance" or suppressor variable should have low correlation with the criterion and substantial correlation with another predictor variable which has in turn substantial correlation with the criterion (compare with page 773).

It has been noted above that in many practical problems the criterion itself constitutes a battery in the sense that several measures of success are obtainable. The combination of these measures of success into a single criterion measure is a problem of importance equal to that of combining prediction measures. Horst (*16*), Edgerton and Kolbe (*10*), and Wherry (*36*) have given both rigorous and approximate solutions to this problem. The rationale of nonstatistical methods of combining criterion elements is given by Brogden and Taylor (*5*).

A somewhat more complex problem arises in the following situation: Suppose we have a battery of scholastic aptitude tests to be used for college placement. The tests measure various aspects of ability and achievement. Suppose also that we have a number of different measures of college success, for example, grades in each of several required freshman courses. It is desired to predict for each entering freshman his probable achievement in each separate course. Here we have a set of several criteria where we wish to predict, not some combination of the criteria, but each criterion measure in turn—in effect a set of multiple prediction problems related only by the fact that the set of predictor variables is the same. When each criterion measure is available for the total sample, Dwyer (*9*) and Mosier (*20*) have shown how the several sets of regression weights may be computed economically by the use of inverse matrices. The methods of Dwyer and Mosier can be used, however, only if the matrix of intercorrelations for the predictors and each criterion is based on the same sample. Unless data for the several criteria are available from the same sample, the intercorrelations of the predictors will differ for each criterion, reflecting differences in the several samples. If, in the example previously cited, we

wished to use the tests of the placement battery to predict grades of engineering students and of liberal arts students, we are faced, on the computational side, with the nasty problem of carrying out, *de novo*, as many full multiple regression solutions as we have criteria to predict. We must, moreover, compute the intercorrelations of the predictor variables separately for each subsample.[9] If the several criteria are psychologically distinct, so that they would be substantially uncorrelated, and if each test in the battery contributes substantially to each criterion, the end may justify the labor.

The final problem in the use of batteries for prediction involves several tests in the battery to predict the best single combination of several criterion measures. Here our problem is not that of differential prediction of course grades, but the prediction of average grade where that average is determined by weighting each course in such a way as to give the most predictable composite grade. This is the problem posed and solved by Hotelling (*17*) in his technique of canonical correlation. It will not be developed here since it goes beyond the scope of this book. We may summarize the problems of linear combination of variables (using least-squares concepts) as shown in Table 15.

Prediction of dichotomous criterion

In many situations the problem of paramount interest is use of a set of measures to predict a dichotomous criterion. This criterion may be success or failure on a job, or in a proposed vocation, or in a course of study. The task of predicting group membership, regardless of how the group may be defined, is one to which the use of the discriminant function is particularly adapted. Application of this technique provides a set of weights for combining the several test scores in such a way as to provide the maximum difference in the average composite score of the two groups, actually, to maximize the ratio of the between-groups variance to the total variance. Also appropriate to this purpose and with the added advantage of preserving the relative proportions in the group is the point biserial correlation coefficient. Selover (*30*), Beal (*2*), Travers (*35*), and Garrett (*12*) give illustrations of the application of the technique, and the latter two present approximation procedures.

Nonlinear additive functions

Thus far we have assumed that the various tests combined in linear com-

[9] If all of the samples have been selected from a common pool, for example, the freshman class, a strong case can be made for using the single matrix of predictor intercorrelations based on that common pool, rather than the predictor correlations based on each sample in turn.

binations ($y = ax_1 + bx_2 \ldots$). Although this is the simplest assumption, and should be used whenever there is no evidence to justify the use of a more complicated one, there is no basis for clinging to it in the face of cogent reasons to the contrary. As we have seen, there are many psychological functions for which it is wholly unreasonable to assume a propor-

TABLE 15

Problems and Sample Techniques for the Linear Combination of Measures

Predictor Variables	Criterion Variables	Conditions	Techniques
1	1		Simple linear regression.
n	1	Each predictor measures same aspect; several used to reduce errors of measurement, i.e., intercorrelations approach reliability.	Addition of standard scores, cf. Holzinger (*13*), Richardson (*26*).
n	1	Each predictor measures a different aspect of criterion or is introduced as a suppression variable.	Multiple regression; Wherry-Doolittle (*38*).
1	n	A set of criterion elements is to be combined into a single composite to be predicted by a single predictor.	Edgerton and Kolbe (*10*), Horst (*14*), Wherry (*36*), Brogden and Taylor (*5*).
n	n	Differential prediction of each criterion variable in turn.	Dwyer (*9*) or Mosier (*20*) if matrix of intercorrelation of predictors same for criterion subsample. Repeated solution of full multiple correlation for each criterion subsample.
n	n	Most predictable combination of the combination.	Hotelling's (*17*) canonical correlation.

tionate increase in the criterion with each unit increase in the measured function throughout the entire range of ability measured. The relationship between visual acuity and automobile driving skill is not a far-fetched or an atypical example. If we consider job success on a relatively simple task, for example, junior clerk-typist, there is little question that success is a fairly linear function of some combination of verbal and numerical ability, within the middle range of these abilities. Below a certain "intelligence" level, however, success in such a job may become substantially zero and further decreases in the measured abilities are probably not paralleled by corresponding decreases in success. Moreover, at the upper levels, ability beyond certain measurable levels finds little opportunity for expression within the limitations of the job and thus is not likely to be accompanied by increased success. In fact, beyond a certain level, increased ability may result in dissatisfaction, loss of interest, diminished motivation, or "ideas beyond

the job," and thus in less than maximum success. It may well be only that we usually deal with a narrow middle range of measured abilities, where the curvilinearity is not marked, that has permitted us to use linear regression as much as we have. Even where the relationship between test and criterion is linear for one logical set of "units," there is no assurance that the logical set of units that we have chosen for our measurements will yield a linear regression. For example, if a criterion is linearly predicted by a test score expressed in units of amount-per-unit time, then the regression of the criterion on a time-per-unit amount will not be linear except as an approximation within a narrow range of scores.

Where the psychological or logical considerations lead us to expect curvilinearity, or where the data show clear indications of a lack of linearity, the possibility of nonlinear combination should be explored provided the number of cases warrants and the findings can be cross-validated. Techniques for dealing with multiple nonlinear regression are available in intermediate statistical texts. One simple technique frequently useful is to modify the scale of measurement in the predictor variables to give a linear plot upon the criterion and to use linear techniques on the modified variables. It is a question of experimental fact, of course, whether or not the variables, rectified in terms of the criterion, will also be linear with respect to each other. If they are not, other methods must be used.

Nonadditive prediction

Up to this point we have discussed these methods of combining the scores of the several elements in a battery by adding scores (or some derived function of the scores) to yield a single composite index. Other types of combination than simple summation are possible. A word of caution should be introduced here, not only with respect to nonadditive combinations but to nonlinear combinations as well. There is no justification for using any more complex hypothesis than simple linear combination if the simpler method will fit the experimental facts as well as the more complex, within the limits of experimental error of the investigation (compare with page 786).

Besides the nonadditive methods of combination suggested above, another type deserves special mention (*31*). This is the use of multiple cutting scores on a series of tests, particularly when the function of the set of tests is merely to divide the examinees into two groups, "pass-fail," "acceptable-unacceptable." Although this is frequently done, its principal justification is one of expediency. A generally better solution is the prediction of the criterion by multiple regression methods and the establishment of a single cutting score on the composite score thus obtained. If, of course,

certain variables are crucial, so that scores below a critical value cannot be made up by higher scores on others, there must be cognizance of this. If, for example, one is predicting success as a member of a submarine crew, men over a certain height may be ineligible, however well they may do on all other variables. In such a case, literal application of linear multiple regression would yield unacceptable results. The use of nonlinear techniques would avoid this difficulty.

In the example given, where there is probably no relation between height and success except at the critical value, the criterion value predicted from all other variables may be multiplied by 1 or 0 depending on whether the individual is within or beyond the acceptable range. Another solution, principally of theoretical interest, is also available. If, as in the example of the relation of visual acuity to driving success, there should be a linear relation between acuity and success within a particular range, the acuity scores may be transformed to another scale, so that all values in the unacceptable range are represented by minus infinity and those beyond the upper cut-off are assigned the same value as the highest acceptable value. In this case, individuals scoring below the critical point will have predicted values of minus infinity, regardless of the weight given to acuity and regardless of how well they may do on other measures. Similarly, individuals scoring above the upper cut-off will not, because of that factor, receive any higher predicted criterion.

In a second type of situation, however, the several measures differ greatly in cost of application, and the examiner is under no obligation to apply all tests to all applicants. There is considerable economy in applying the most economical measure first and eliminating all those who fall below the critical score, even though some of those, by virtue of exceptionally high scores on later tests, might score above the passing point for composite score on the total battery. Obviously, if such a procedure is followed, the first test given should, to the extent possible, be not only the most economical, but also the one which would have the highest beta weight in the final regression. Moreover, at each step in the process, elimination should be based, not on the last test alone, but on the composite of all administered up to that point. When multiple cutting scores are set, they should be set low enough that the number passing any one test is considerably in excess of the number expected to pass the entire set. Otherwise the number finally passing all hurdles will not meet selection ratio requirements.

Differential weighting

It is frequently possible to use the same battery to predict several criteria. One example of this is the Differential Aptitude Tests where the

same set of measures is used to obtain predictions of success in a variety of school subjects. Another example is afforded by the Army Classification Battery (*1*) where varying combinations of ten tests are used to predict success in a wide variety of jobs. Crawford and Burnham at Yale have used the College Entrance Examination Board Scholastic Aptitude Test battery to predict success in each of several collegiate curriculums (*6*). The Army Air Forces used a single basic battery and differential weights to predict success as pilot, navigator, and bombardier (*11*).

To the extent that the several criteria are statistically independent and the battery contains the major determinants of success in each criterion, differential weighting is a technique which deserves much wider use than it has received. It suffers two limitations. If the battery is to include the major determinants of each of several criteria, it is likely to be so long as to be administratively unfeasible, or to be composed of such brief measures as to base the predictions on unreliable measures. Moreover, for the prediction of any one variable, a number of the elements in the battery will prove to be dead wood, even though important in the prediction of other criteria.[10] Unless predictions of every criterion are important for comparison within the individual, for example, in guidance or placement, the use of selected batteries to predict selected criteria is likely to prove more feasible.[11] Moreover, unless two criteria approach statistical independence, the value of criterion A predicted from an individual's test scores will be substantially the same as the predicted value of criterion B. In the prediction of multiple criteria we are usually concerned with the differences between them; that is, we are interested in such questions as these: Will John do equally well in engineering as in liberal arts or substantially better in one than the other? Does Bill have higher interests as accountant than as purchasing officer? As was pointed out earlier, the reliability of the differences between regressed scores depends both on the reliability of the individual predictions and on the correlation between the predicted scores. Unless the criteria are relatively uncorrelated, we are apt to read significance into differences in predicted behavior which are based largely on chance error rather than upon stable differences characteristic of the individual.

A special case of differential weighting for the prediction of several criteria which deserves special mention, though this is not the place for extended treatment, is the prediction of factor scores from a battery which

[10] In differential *description,* where we are exploring a student's or counselee's abilities prior to selecting specific areas for prediction purposes, a battery including many diverse ability measures is essential.

[11] It may often be desirable, however, to administer all tests in such a battery at least once to ascertain which tests are predictive of each criterion. Empirical evaluation frequently differs considerably from arm-chair judgments of probably appropriate tests.

has been subjected to multiple factor analysis. In this case the criteria are not given externally, but are derived as hypothetical variables from the results of the analysis.[12] If the results of the analysis are expressed in orthogonal factors, then the question of independence of the criteria is disposed of by definition. As more factors are identified and verified and as more is learned about the relationship of such factors to the socially significant variables which constitute the usual external criterion—school success, job performance, and so on—the prediction from factor scores will assume greater importance. Whenever a number of criteria are to be predicted for comparison among the same group of subjects, the careful choice of a battery and the use of differential prediction provide great economies in the over-all testing, even though each criterion could be predicted as efficiently with fewer than the total number of predictors used.

The foregoing analysis of batteries in terms of factorial composition is far more applicable to aptitude than to achievement batteries. In the latter case, the function of the battery goes beyond prediction of an external criterion and is concerned with information on achievement in curriculum units which are defined in terms other than their factorial components. Until much more is learned from the application of multiple factor techniques to achievement testing problems, we can do little toward the building of achievement batteries with very low intercorrelations among the separate tests. So long as we are concerned with achievement in specific, defined curriculum units, and so long as those curriculum units comprise complexes of highly overlapping knowledges, skills, and abilities, then intercorrelation among the components appears unavoidable. The possibility of a core test covering the common aspects of a number of fields, supplemented by subject-matter tests covering only those knowledges and skills unique to each subject matter, might well be explored. It seems likely, however, that the complexities of present subject-matter organization make such a solution unattainable. Another possible line of inquiry is the application of factorial or other similar approaches to achievement testing results to identify the areas of overlapping skills among apparently parallel and discrete subject-matter areas so that undesirable duplications may be eliminated and desirable duplication attained on a directed, rather than a sporadic, basis. Such research is, of course, beyond the scope of this work.

Profiles

Because profiles are simple to construct and are superficially easy to

[12] A word of caution is in order. So-called "factorial validity" is not a substitute for experimental validity as considered heretofore. Prediction of external, socially important criteria is the prime concern of measurement; prediction of factors is a technique for describing a test. No matter how factorially pure a test is, we still judge its usefulness on the basis of prediction of socially meaningful criteria, external to the test itself.

interpret, they constitute one of the most popular methods of summarizing the results of multiple measurement. They have the added advantage that the graphic presentation enables one to "picture the total set of test scores and their interrelations at a glance." Here, more than in any other aspect of test interpretation, do we need to beware of seeming simplicity.[13] By failing to question the reliability of differences between scores, and by relying on the judgment of the interpreter to make the over-all summary, we ignore the possible unreliability and invalidity of the over-all summation which would be instantly revealed if less "simple" methods were used. This comment, supported by more detailed consideration later, should not necessarily lead to the abandonment of the profile, but to an awareness of the precautions which should be followed in its use.

Elementary principles of profile construction

In its usual form, a profile is a graphic representation of a set of test scores for a single individual in which the tests are represented by ordinates spaced along the horizontal base line and the magnitude of each score is represented by plotting the point at the appropriate height on that ordinate. In order to aid the eye in locating the points thus plotted, it is customary to join the points by lines, leading to the more or less "jagged" picture that gives the technique its name. It should be remembered that the line thus drawn is not a graph in the usual sense. We are accustomed to dealing with graphs of continuous functions, even though only a few points may be experimentally determined. In such graphs, the lines have meaning as representing values associated with values intermediate between those plotted. This is not true of ordinary test profiles. What would be the meaning of the point x in the sample profile illustrated in Figure 70? Some of the objections to profiles may be overcome by plotting profiles without connecting the profile points. Two methods of doing so are illustrated in Figure 71. These profiles avoid the false assumptions inherent in the connecting of score points on the tests, as well as being less subject to configurational misinterpretation. If the plotting is made on the basis of an identified group, with the median of the group shown as the line from which the student deviates, comparison of the student with the group, as well as impressions of the individual's relative strengths and weaknesses, is fostered. How much configurational misinterpretation remains with such profiles is difficult to estimate.

Since raw test scores may vary considerably in meaning, it is obvious that raw scores cannot be used in plotting the profile. For the vocabulary test, raw scores may range from 50 to 119, while those for arithmetic may vary from 7 to 36. If raw scores were used, it is apparent that the arbitrary nature

[13] Compare with pp. 803–4.

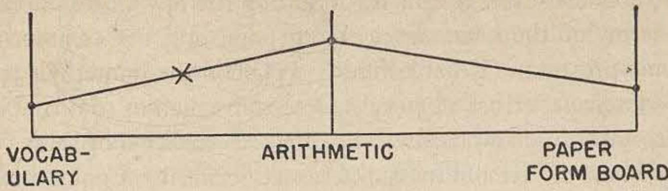


FIG. 70.—Hypothetical profile showing error involved in connecting profile points

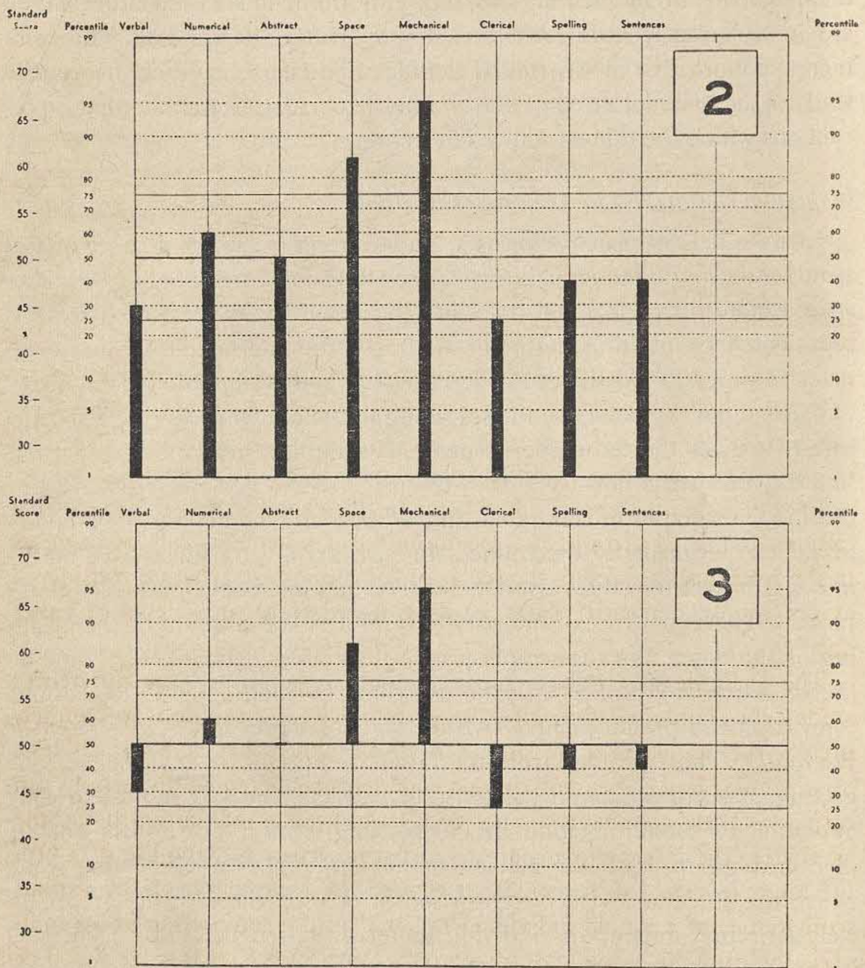


FIG. 71.—Two types of profiles not involving connected profile points. (Figures 2 and 3 from pages D-4 of *Manual for the Differential Aptitude Tests,* published by the Psychological Corporation.)

of the scores makes it certain that no one can ever appear "better on arithmetic than on vocabulary." Before a profile can be plotted, then, it is clearly necessary to reduce the scores to comparable values (see pages 750–

60). This may be done, of course, by scaling the raw scores on each ordinate in terms of the common reference point and the common unit of measurement so that it is not necessary to obtain the numerical equivalent for each raw score before plotting it. The conversion to comparable scores is thus accomplished graphically. The essential condition of these comparable scores is that each test be measured from the same basic point of reference and in the same unit of measurement. Various assumptions or definitions may be involved in accomplishing this. The most common is to use either standard scores or percentile ranks scaled proportional to standard score distances.[14] Note that when this is done, *the standard scores or percentile ranks must be based on the same or strictly comparable populations.*

John Smith		Test		
		Learning Ability	Mechanical Aptitude	Clerical Aptitude
John Smith's	Raw Score	78	47	213
John Smith's	Percentile Rank	48	37	72
	Normative Group	*College Freshmen, X-College*	*High School Senior Boys, Shopwork*	*Employed Women in Clerical Jobs*

In which area is the individual best? In this example, possibly somewhat exaggerated, it is obvious that no meaningful statements can be made as to the relationships among the scores of the individual, whether considered as raw scores, percentile ranks, or as standard score equivalents of either one.

The example given above has deliberately exaggerated the differences among the standardization groups. Nevertheless, if a profile is to be interpreted, the standardization groups must be either the same for all tests or so similar that the assumption of equal means and equal dispersions can be made. These comments are not peculiar to profiles, of course; they apply as well to the interpretation of any test score. However, the use of a profile tends to establish a mental set toward the comparison of the various score values of a single individual rather than the comparison of an individual with other individuals of a group. With such a set the likelihood of ignoring the differences in the groups is greatly increased.

Another type of unit in which profiles are often plotted is the public school grade equivalent. Even where equal standard score units, or school-

[14] In the first instance, the assumption of equal score units throughout the range is implied; in the second, normality of distribution is assumed. In either case, comparison of two sets of scores is possible only if the means and sigmas respectively of the two are equated.

grade equivalents are used, the results must be interpreted with full knowledge of the factors contributing to the results. As an instance of this, consider the hypothetical profile of a superior student in the eighth grade, shown in Figure 72. The student is clearly superior—approximately two full grades advanced—in English, social studies, and general science. In

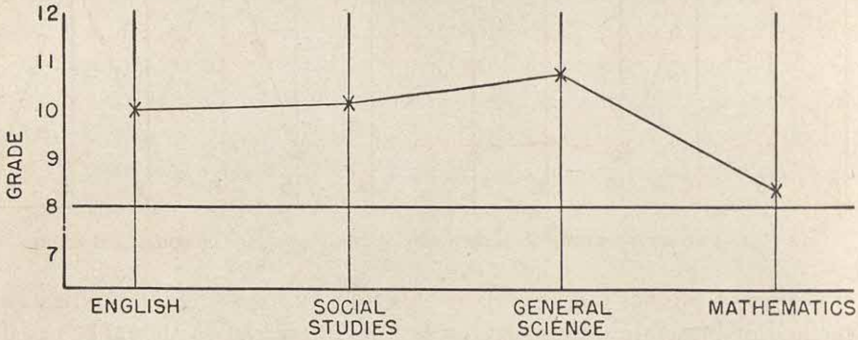


FIG. 72.—Hypothetical profile of a superior student in the eighth grade

mathematics, however, his showing is by no means as good; it is true he is ahead of his class, but by only a half-grade instead of the 1½–2½ grades in the other fields. Does this reveal a weakness for mathematics? Possibly; but far more likely it reveals that mathematics is far more closely integrated into the school sequence than are the other fields. The superior student may and does read widely in literature, contemporary affairs, and general science. In class he may receive advanced or enriched assignments. But the formal mathematics characteristic of the ninth-grade curriculum is likely to be learned only in a ninth-grade mathematics class. If the method of solving of quadratic equations is introduced in ninth-grade mathematics, how is the eighth-grade student, however advanced he may be, likely to come by that knowledge?

A problem which has not been given the consideration it deserves is the question of the arrangement of the horizontal ordinates. In which order should a set of tests be arranged for plotting? Since all of the scores are presented, and since the lines between plotted points are meaningless, it may be that order on the base line is wholly immaterial. On the other hand, the interpretation of the "pattern" of the profile is usually made as a psychological judgment, based, not only on the numerical values of the scores, but on their total perceptual configuration. To the extent that this latter factor enters, order is important. Consider the impression made by the two profiles shown in Figure 73, each based on the same set of test scores. Which is the easier to interpret; which the most open to misinterpretation? To the writer's knowledge, no investigation of these problems has been

made directly, although the field of visual pattern perception should offer useful hypotheses for testing. Without awaiting research answers, the prac-

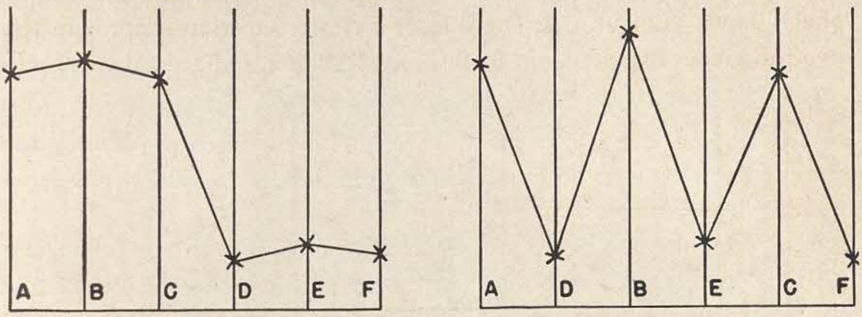


FIG. 73.—Two arrangements of same profile showing possible configurational errors

tical user of profiles can avoid possible errors by regarding the profile as a method of presenting the actual test results. Interpretation should be based on test scores aided by the profile, not on the profile aided by the scores.

Types of profiles

Thus far, we have considered, to illustrate the general nature of profiles, only one type. In these profiles, scores on several different tests obtained by a single individual at a single time have been plotted to give a static cross section of his scores. In the longitudinal profile, scores on the same test or set of tests at different times are plotted to give a picture of the growth of the individual in the functions measured. In these profiles, the question of ordering the test ordinates does not arise (except within each set of scores) since the time sequence determines the ordering. The spelling, arithmetic, and reading ability of the pupil can be charted from the fourth grade through the eighth. If the number of tests at each grade becomes greater than two or three, the chart becomes so complex that interpretation is extremely difficult. Moreover, the problems of finding a set of vertical scale values that will accommodate not only several tests, but widely different age or grade groups, and will at the same time meet the necessary conditions that all tests plotted be standardized (for score purposes) on comparable samples, becomes one calling for extensive, coordinated reseach in the standardization.

If these difficulties are surmounted and proper precautions are taken in interpretation, graphic presentation of data in a profile does provide a means of presenting a number of facts in compact form suitable for many practical applications. When statistical methods of combining or comparing scores are feasible, however, they are much to be preferred.

Reliability of profiles

The concept of the reliability of a test score is thoroughly grounded among test users. The notion that profiles have varying degrees of reliability is not so well established. Obviously, each point plotted is a test score, and as such is subject to the same errors of measurement as that same score expressed in numbers (see page 778 for one method of insuring recognition of this). Certainly no greater confidence can be given the scores when plotted in a profile than is given those same values considered as individual test scores.

Moreover, in profiles we are concerned, not only with the magnitude of the scores, but also with those differences among them which constitute the essence of the "score pattern." We go beyond the interpretative statements: "John is very high in verbal ability, moderately high in the number factor, and only average in spatial ability." We make in addition other interpretative statements: "John is *higher* in verbal ability than in number and spatial abilities; he is *higher* in number than in spatial; and lower in spatial than in either of the others. Therefore, he does not have the score pattern of an engineering student." (Similar comparative conclusions are drawn from achievement test scores, whether for an individual, a classroom, or a school system.) Yet the unreliability of the differences on which the foregoing interpretation hinges may be such that John could actually be equal in all three, or higher in spatial than in number. Again, the very concreteness of the graphic pattern gives it an appearance of accuracy that is wholly spurious. After all, there *are* the scores, not only in black and white, but as points on a graph.

As pointed out above, the interpretation of a profile usually depends, not only upon the scores, but upon their interrelationships, that is, upon the differences among them. But the reliability coefficient (and the interpretation) for the difference between scores of a single individual is a function, not only of the reliability of the two tests, but also of the correlation between them. The formula for this reliability is:

$$r_{(z_1-z_2)(z_1'-z_2')} = \frac{r_{11'} + r_{22'} - r_{12'} - r_{1'2}}{2\sqrt{(1-r_{12})(1-r_{1'2'})}},$$

where 1′ and 2′ refer to second observations on traits 1 and 2. A general aggregate formula for the stability of a profile is given by:

$$R = \frac{\bar{r}_{ii'} - \bar{r}_{ij'}}{\sqrt{(1-\bar{r}_{ij})(1-\bar{r}_{i'j'})}},$$

where i and j are tests, i' and j' are second measurements of the traits meas-

ured by i and j and $\bar{r}$ denotes the mean correlation coefficient. If tests i and i' are parallel this reduces to $\frac{\bar{r}_{ii} - \bar{r}_{ij}}{1 - \bar{r}_{ij}}$.

Interpretation of profiles

As with single test scores, or any other measurement, interpretation of a set of scores involves comparison with the performance of known groups. Even though each test score is expressed in relation to some standardization group for the purpose of rendering the scores on each test in the set comparable with one another, there remains the problem of interpreting the profile or pattern. Suppose we have the profile shown in Figure 74. How shall we interpret it beyond a bare recital of the numerical values of the test scores? Obviously, we cannot. And just as obviously, we cannot interpret it merely by replacing test numbers with test names.[15] The reader who has come this far realizes by now, if not before he opened the cover, that the meaning of a test score cannot be inferred from the label which the test maker or publisher decides to attach to the test. If the tests are familiar to us, so that we know the performance of the standardizing group, and perhaps of other meaningful groups, then each score may be given additional meanings, but we still are not in a position to *predict* the past or future behavior of John Jones from the profile alone. If, besides, the tests are so familiar that we know the correlates of high and low scores on the tests, then we can make predictions concerning these correlates for *each* test considered singly or for the tests considered as a battery. We have still gained nothing in our interpretation by plotting the scores in a profile.

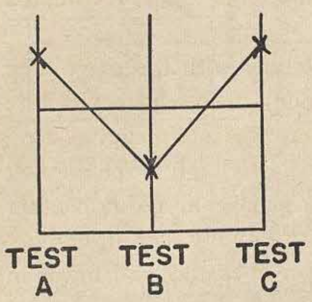


FIG. 74.—Schematic profile omitting all interpretational data.

In considering the effectiveness of interpretation from profiles, many of our questions relate to any normative score, whether profile or not. It is worth noting, however, that the use of a profile does not solve the questions, and should not lead to ignoring them. Let us add to our information shown in Figure 74, however, that profiles representing the mean scores for tests A, B, and C for three occupational groups (defined much more fully as

[15] Test scores do not differ from physical measurements in this respect. A temperature of 96°F. is hot if compared with the mean May temperature in New York, but low if read in Flagstaff, Arizona, in July, and a danger signal if it is the body temperature of an adult male human. Similarly, the meaning of inch varies depending on whether it is compared with nose-lengths or height.

to age, sex, length of employment, educational level, and so forth) are as shown in Figure 75.

Now certain meaningful statements can be made about the profiles of file clerks (female?) law school students (college graduates?), engineers (ten years on the job?). The questions in parentheses, insofar as the answers

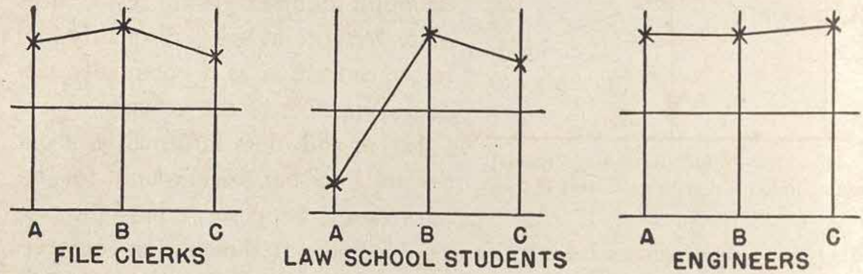


FIG. 75.—Hypothetical profiles representing mean scores for Tests A, B, and C, for three occupational groups—file clerks, law school students, and engineers.

affect the group's profile can and should be answered. We can now say whether or not John's profile corresponds to the profile (pattern of test scores) of the average file clerk. But if it does, is the average file clerk successful? Is the individual law school student more or less successful as his profile lies wholly above or wholly below the average of the group?

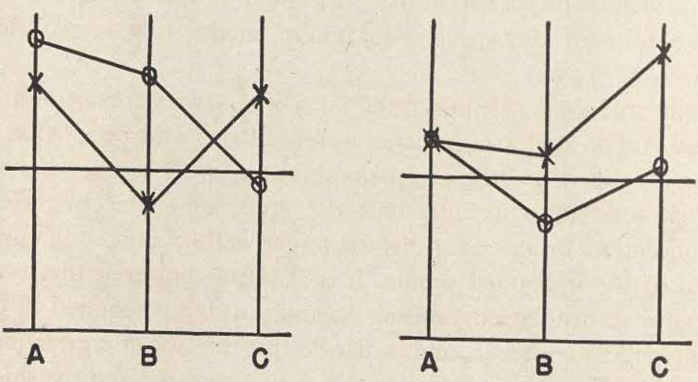


FIG. 76.—Sample profile of individuals superimposed on group profiles

What shall we say of the individual whose profile, when compared with that for the group in which we are interested, shows several points above the group mean, but one point conspicuously below it? To what extent, then, does superiority in one or more tests compensate for marked deficiency in another? Consider the profiles shown in Figure 76. Here the individual's scores are plotted as X's, the mean scores of the comparison group are

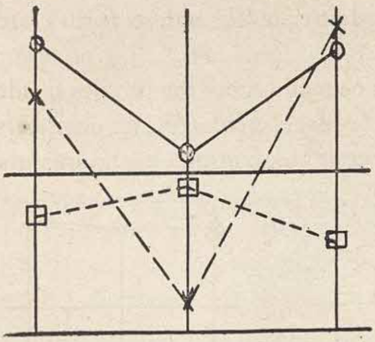


FIG. 77.—Profile illustrating comparison of individual with successful and unsuccessful groups.

shown as O's. The first profile illustrates the question of compensability raised above. The second raises a new question. Although the individual's profile corresponds closely to that of the norm group except for test C, does his *superiority* in test C disqualify him for consideration as a potentially successful member of the group?

Let us add more information about the tests to our background for interpretation. Suppose we have the profiles, not only of successful but also of unsuccessful freshmen in an engineering curriculum. In Figure 77 the successful group means are plotted as O's, the unsuccessful means plotted as □'s, and the individual's scores plotted as X's.

Since test two does not serve to differentiate successful from unsuccessful engineering students, what significance shall we attach to the fact that John's score is well below the value of even the unsuccessful group? Shall we say that the score on that test is irrelevant to success in engineering school and, therefore, that test may be ignored? Or shall we make the equally plausible interpretation that since John on this test falls far below the mean of even the unsuccessful group, engineering is probably not for him?

The dilemma above would become worse if test two were one where both successful and unsuccessful students score well above the population mean. It is true that there is little correlation between this test score and success when that correlation includes only the group of engineering students. If we concentrate on this fact, this test might well be ignored in our interpretation of the individual profile. It is also true, however, that even the unsuccessful student of engineering is, and possibly is required to be, far above the population mean, and it may be that the lack of correlation with success results from the selective factors which have reduced the spread of differences on this variable. In this case an average score on the test indicates that the individual falls completely outside the group of engineering students as a whole, much less the subgroup of successful students.

The foregoing questions point toward another which rises to vex us in reading a profile: What quantitative answer can we give as to the limits of tolerance within which the individual's profile must agree with the criterion profile to which he is being compared? Obviously, the individual's score on each test contains an error factor because of the unreliability of the test.

This error can be estimated by the standard error of measurement, and the individual's profile more properly drawn, not as a line, but as a band, whose position and width at each test ordinate are given by the X ± S. E. measures. The individual's true score can be expected to be within this band with odds of 2-1. However, the profile of the normative group is not known exactly, but with a similar range of tolerance. Although this range of tolerance, like that of the individual scores, depends on the reliability of the individual scores on which it is based, the accuracy of the criterion score for all practical purposes may be expressed as the standard error of the mean (or other statistic) used to define the criterion profile. In actuality, then, we have not two lines to be compared, such as those shown in Figure 75, but two bands to be judged in terms of the probability that the individual's score (which probably lies on one band) is below, at, or above the critical score for the criterion group (which probably lies in the other band).

With these difficulties plaguing us when we attempt a quantitatively precise interpretation of a profile, it is small wonder that the average test user chooses to ignore them (though he cannot dispose of them) and turns to simpler, rule-of-thumb methods. One of these, the use of multiple cutting scores, we have considered above. If the individual's profile is at or above the criterion profile for all tests, there is little question that he meets the specifications of the criterion group—in the tests specified.[16] If he falls below the cutting score on any one test, he is judged as failing to meet the specifications. If the tests specified are valid predictors of membership in the criterion group and if the cutting scores have been appropriately chosen, such conclusions, though gross, are probably justified. They have the disadvantage of being much more restrictive than the data require. Moreover, their use precludes the possibility of compensating for a low score on one test by a higher score on another, and many individuals are disqualified unnecessarily.

Perhaps the most used "method" of interpreting a profile is one which may be termed, for want of a better name, "impressionistic evaluation." The profile is scanned by the counselor who then reaches an over-all judgment in the light of the test scores, their interrelations as shown by the profile pattern, his apperceptive mass of information regarding the tests, and the extent to which they characterize various types of behavior of the sort to be predicted. The counselor is able to consider in this over-all judgment, not only what he knows about the test scores, and the tests, but all of the other information about the individual as well. This totality

[16] In certain situations it may be necessary to set an upper critical score as well, above which the individual should be eliminated. This most often occurs with low-level or routine tasks.

of information is then synthesized into a composite picture through a judgmental process which attributes to each item a weight determined by the counselor's subjective opinion.

The use of impressionistic evaluation for profiles has here the same advantages and the same disadvantages that it has in the interpretation of any other type of test data. It is not unique to the evaluation of profiles, and the use of the profile in the process neither particularly enhances nor detracts from its effectiveness. It permits the consideration of qualitative data which more exact methods often ignore because there is no simple way of including them. Far more complex situations can be evaluated in this way, since multiple regression equations involving ten or twenty variables, though theoretically possible, are computationally awkward, to say the least. Finally, in the eyes of the advocates of the method, it permits an evaluation of the total person, which considers the interrelation among attributes as well as their absolute value and permits consideration of such interrelationships, not only in terms of their statistical intercorrelations in the population of persons, but in terms of their mutual interaction within the individual. At the same time, it attempts by subjective and variable means to approximate the results of the multiple regression problem—the best prediction from a combination of several measures—without a basis for making the approximations very close and usually without a way of checking on its effectiveness. Brogden's summary (*4*) directed toward the use of profiles in clinical diagnosis is appropriate here:

> Profiles are in general subject to misleading interpretation. It is very often implied or openly stated that it is the degree of adherence to the pattern shown by the profile which is best used to determine membership in a diagnostic category. While this may occasionally be true it has never to the writer's knowledge been demonstrated. It is entirely unsatisfactory merely to select cases according to their subjectively estimated agreement with the profile of the mean standard scores of a diagnostic category and to demonstrate that such selection is more efficient than would be explained by chance. Significance of mean differences in individual test scores between the given category and the standardization population for 3 out of 10 tests (or even one test, for that matter) would result in a profile of mean differences of promising appearance. Selection of cases by degree of agreement with the profile of means could be obtained to a degree unexplainable by chance even though the remaining tests were completely unrelated either to the tests or to the diagnosis. Such a profile would be much less efficient in differentiating than the 3 valid tests substituted in a regression equation. Suppose on four tests having means of ten and SD's of three in the standardization sample, mean scores of 10, 14, 10, 12 were obtained. With zero intercorrelations of the tests in both the standardization and the diagnostic category an individual having scores of 6, 10, 6, 8 would be no more likely to belong in the diagnostic category than would an individual having scores of 14, 10, 14, 8.

> Numerous other examples could be cited. Use of profiles is justifiable only when it has been demonstrated that patterns occur that are not explainable in terms of summed scores.

In summary, profiles represent one way of presenting and summarizing the results of multiple measurement. Profiles can present graphically either a cross section of the individual's pattern of traits at a particular moment of time—a number of different tests administered at the same time—or a longitudinal pattern in which a small number of tests is repeated at various stages in the growth of the individual. Profiles of this latter sort are not strictly profiles, but growth curves.

Since profiles are merely the graphical representation of fallible test scores, they must be interpreted in terms of the reliability or unreliability of the tests, and in terms of the intercorrelations among the tests represented. Since the meaning of a profile depends upon relationships among the measures of the type A > B, we must be concerned with the reliability not only of A and B but of the difference A–B. If test results are to be plotted in profile form, it is imperative that the scores of all of the tests be reduced to a common origin and a common unit of measurement, since the purpose of profiles and the method of presentation force comparison among the various test scores of the same individual. Common origin and units are given by standard scores, normalized scores, or scaled percentile ranks only if the standardization groups are identical for all tests or represent large, unbiased samples from the same population. These conditions are occasionally met, but not as frequently as profiles are used in practice.

Unless we are to fall into the error of attributing reality to the *name* assigned the tests, the evaluation of an individual's profile must be made in terms of its meaning for the prediction of the outside, meaningful behavior in which we are interested—success or failure in a particular line of activity, satisfaction or dissatisfaction with a certain type of work, satisfactory or unsatisfactory emotional adjustment under foreseeable stresses and strains. To do this, we must say that this profile possesses the essential characteristics of those who are or are not successful, adjusted, or satisfied in various degrees. This comparison of the individual profile with that of the criterion profile may take several forms: (*a*) comparison of peaks and valleys, formalized by the rank difference correlation between individual scores and criterion scores on the several tests (thus ignoring differences in level); (*b*) comparison with a stencil pattern of the criterion profile or profile band wide enough to allow for the limits of tolerance; (*c*) use of multiple cutting scores; and (*d*) impressionistic evaluation. Careful consideration of all of these leads to the conclusion that none of them is as effective as the use of multiple correlation to provide, for each criterion to be predicted,

a composite score which is then subject to whatever further interpretation may seem desirable. Moreover, the use of facilitating tables and other computational shortcuts makes the computation of such composite scores for the criterion in which we are interested little, if any, more tedious than the profile plot. As to feasibility, unless the data required for multiple regression are available, namely relationship of each test with the criterion and with each other test, we have no basis on which to make any type of appraisal of the meaning of the set of test scores. The principal advantage of the profile is its apparent simplicity in presenting the results of testing to the individual tested or to others whose appreciation of test scores is wholly at the nontechnical level. In such presentation of profiles, however, the test technician has an obligation to base his interpretation of the meaning of the scores on more scientific methods and more extensive data than are given by intuitive evaluation of the arrangement of scores on a grid.

Selected References

1. Adjutant General's Office, Department of the Army. *Army Classification Battery Manual.* DA-AGO-PRT 734. Washington: The Department, 1949.
2. Beal, G. "Approximate Methods in Calculating Discriminant Functions," *Psychometrika,* **10**: 205–18, 1945.
3. Bennett, George K., and Doppelt, J. E. "The Evaluation of Pairs of Tests for Guidance Use," *Educational and Psychological Measurement,* **8**: 319–25, 1948.
4. Brogden, H. E. *The Relationship of Army Individual Test Subscores and Other Mental Ability Tests to Diagnosis of Mental Disorder.* ("PRS Report," No. 724.) Washington: Adjutant General's Office, War Department, 1946. Mimeo. 33 pp.
5. Brogden, H. E., and Taylor, E. K. "The Dollar Criterion—Applying the Cost Accounting Concept to Criterion Construction," *Personnel Psychology,* **3**: 133–54, 1950.
6. Crawford, A. B., and Burnham, P. S. *Forecasting College Achievement.* New Haven, Conn.: Yale University Press, 1946. 292 pp.
7. Derflinger, G. W. "Prediction of College Success: A Summary of Recent Findings," *Journal of the American Association of Collegiate Registrars,* **19**: 68–78, 1943.
8. Dwyer, Paul S. "The Determination of the Factor Loadings of a Given Test from the Known Factor Loadings of Other Tests," *Psychometrika,* **2**: 173–78, 1937.
9. ———. "The Simultaneous Computation of Groups of Regression Equations and Associated Multiple Correlation Coefficients," *Annals of Mathematical Statistics,* **8**: 224–31, 1937.
10. Edgerton, H. A., and Kolbe, L. "The Method of Minimum Variation for the Combination of Criteria," *Psychometrika,* **1**: 183–87, 1936.
11. Flanagan, John. *Personnel Research in the Air Force,* Vol. I. Washington: Government Printing Office, 1947.
12. Garrett, Harry E. "The Discriminant Function and Its Use in Psychology," *Psychometrika,* **8**: 65–79, 1943.
13. Holzinger, Karl. "Factoring Test Scores and Implications for the Method of Averages," *Psychometrika,* **9**: 155–67, 1944.
14. Horst, Paul. "Mathematical Contributions," *The Prediction of Personal Adjustment.* ("Social Science Research Council Bulletin" 48.) New York: The Council, 1941. Pp. 403–47.
15. ———. "Measuring Complex Attitudes," *Journal of Social Psychology,* **6**: 369–74, 1935.
16. ———. "Obtaining a Composite Measure from a Number of Different Measures of the Same Attribute," *Psychometrika,* **1**: 53–60, 1936.
17. Hotelling, Harold. "Relations between Two Sets of Variates," *Biometrika,* **28**: 321–77, 1936.

18. KELLEY, E. LOWELL, and FISKE, DONALD W. "The Prediction of Success in the VA Training Program in Clinical Psychology," *American Psychologist,* 5: 395–406, 1950.
19. MCNEMAR, QUINN. *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin Co., 1942.
20. MOSIER, CHARLES I. "Determining a Simple Structure When the Loadings of Certain Tests Are Known," *Psychometrika,* 4: 149–62, 1939.
21. ———. *Evaluating Rural Housing.* Gainesville, Fla.: University of Florida, 1942.
22. ———. "Machine Methods of Scaling by the Methods of Reciprocal Averages," *Proceedings of the Research Forum, Endicott, N.Y., August 26–30, 1946.* New York: International Business Machines Corp.
23. ———. "Measurement in Rural Housing, A Preliminary Report," *Educational and Psychological Measurement,* 2: 139–52, 1942.
24. ———. "On the Reliability of a Weighted Composite," *Psychometrika,* 8: 161–68, 1943.
25. RAPAPORT, DAVID, with collaboration of Schafer, Roy, and Gil, Merton. *Diagnostic Testing of Intelligence and Concept Formation,* Vol. I of *Manual of Diagnostic Psychological Testing.* ("Josiah Macy, Jr., Foundation Review Series," Vol. 2, No. 2.) 1944. 239 pp.
26. RICHARDSON, M. W. "The Combination of Measures," *The Prediction of Personal Adjustment,* by Paul Horst, *et al.* ("Social Science Research Council Bulletin" 48.) New York: The Council, 1941. Pp. 377–401.
27. ———. "The Logic of Age Scales," *Educational and Psychological Measurement,* 1: 25–34, 1941.
28. SARBIN, T. R. "A Contribution to the Study of Actuarial and Individual Methods of Prediction," *American Journal of Sociology,* 48: 593–602, 1943.
29. SEGEL, DAVID. *Differential Diagnosis.* Baltimore: Warwick & York, 1934.
30. SELOVER, R. B. "A Study of Sophomore Testing Program at the University of Minnesota," *Journal of Applied Psychology,* 26: 296–307, 1942.
31. SPRINGBETT, B. M. "A Method of Obtaining Critical Scores from Successive Residues," *Bulletin of the Canadian Psychological Association,* 4: 28–30, 1944.
32. THOMSON, G. H. "Weighting for Battery Reliability and Prediction," *British Journal of Psychology,* 30: 357–66, 1940.
33. THURSTONE, L. L. "The Mental Age Concept," *Psychological Review,* 33: 268–78, 1926.
34. TOOPS, H. A. "The L-Method," *Psychometrika,* 6: 249–66, 1941.
35. TRAVERS, R. M. W. "The Use of a Discriminant Function in the Treatment of Psychological Group Differences," *Psychometrika,* 4: 25–32, 1939.
36. WHERRY, R. J. "An Approximation Method for Obtaining a Maximized Multiple Criterion," *Psychometrika,* 5: 109–15, 1940.
37. ———. "Test Selection and Suppressor Variables," *Psychometrika,* 11: 239–49, 1946.
38. ———. "The Wherry-Doolittle Test Selection Method," *Occupational Counseling Techniques,* by William H. Stead, Carroll L. Shartle, *et al.* New York: American Book Co., 1940. Appendix V (pp. 245–52).
39. WHERRY, R. J., and GAYLORD, R. H. "Test Selection with Integral Gross Score Weights," *Psychometrika,* 11: 173–83, 1946.
40. WRIGHT, RUTH. "A Factor Analysis of the Original Stanford-Binet Scale," *Psychometrika,* 4: 209–20, 1939.
41. BROGDEN, H. E. "An Approach to the Problem of Differential Prediction," *Psychometrika,* 11: 139–54, 1946.
42. THORNDIKE, R. L. "The Problem of Classification of Personnel," *Psychometrika,* 15: 215–36, 1950.

INDEX

Index

THE AMERICAN COUNCIL ON EDUCATION

The American Council on Education is a *council* of national educational associations; organizations having related interests; approved universities, colleges, and technological schools; state departments of education; city school systems; selected private secondary schools; and selected educational departments of business and industrial companies. It is a center of cooperation and coordination whose influence has been apparent in the shaping of American educational policies as well as in the formulation of American educational practices during the past thirty-three years. Many leaders in American education and public life serve on the commissions and committees through which the Council operates.